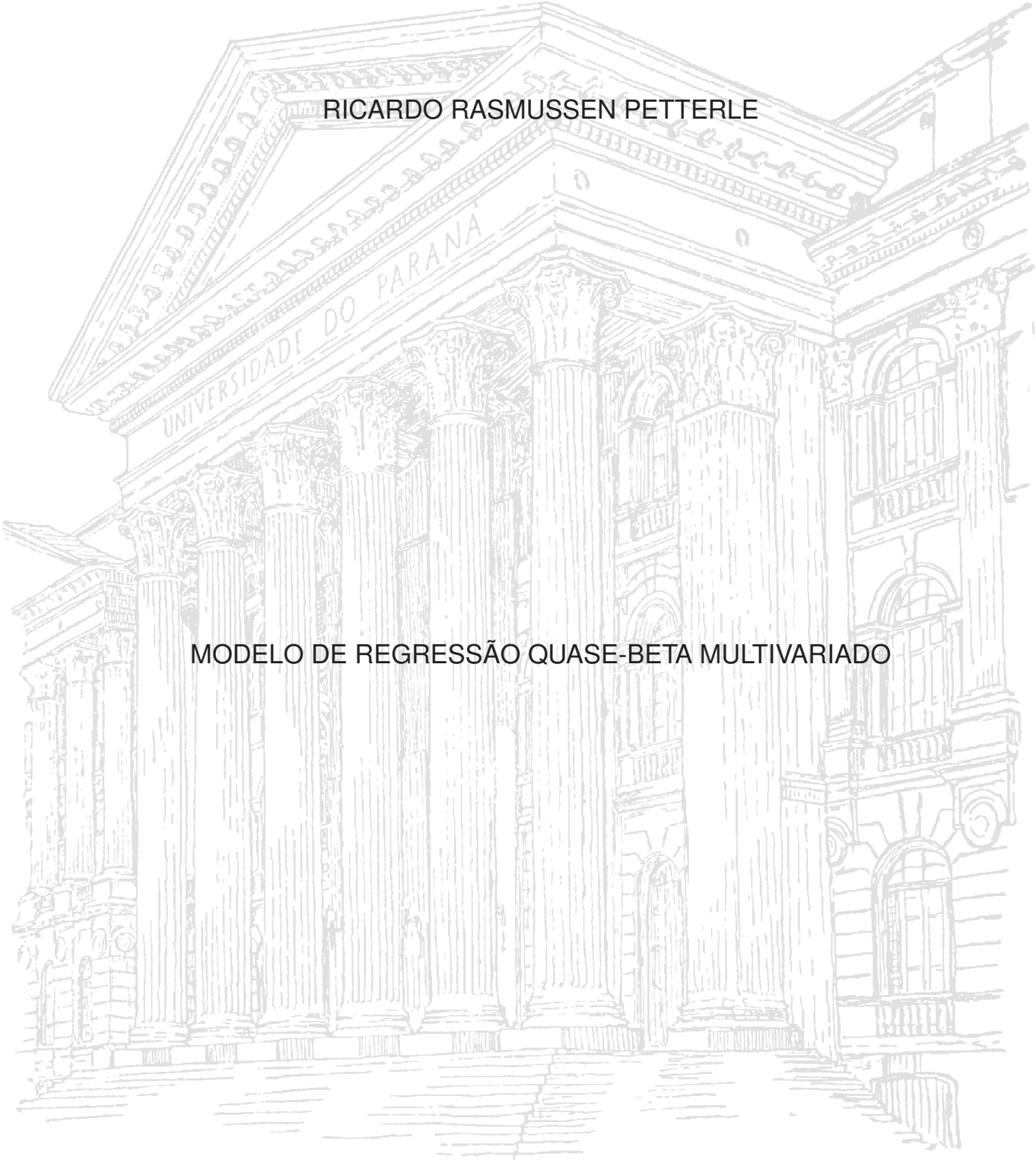
UNIVERSIDADE FEDERAL DO PARANÁ

RICARDO RASMUSSEN PETTERLE



MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO

CURITIBA

2018

RICARDO RASMUSSEN PETTERLE

MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO

Dissertação apresentada ao Curso de Pós-Graduação em Engenharia de Produção, Área de Concentração em Pesquisa Operacional, Departamento de Engenharia de Produção, Setor de Tecnologia, Universidade Federal do Paraná, como parte das exigências para a obtenção do título de Mestre em Engenharia de Produção.

Orientador: Prof. Dr. Cassius Tadeu Scarpin

Coorientador: Prof. Dr. Wagner Hugo Bonat

CURITIBA

2018

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

P499m

Petterle, Ricardo Rasmussen
Modelo de regressão quase-beta multivariado [recurso eletrônico] / Ricardo Rasmussen Petterle. – Curitiba, 2018.

Dissertação - Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-Graduação em Engenharia de Produção, 2018.

Orientador: Cassius Tadeu Scarpin – Coorientador: Wagner Hugo Bonat.

1. Análise de regressão. 2. Estatística. 3. Métodos de redes múltiplas (Análise numérica). 4. Métodos de simulação. 5. Algorítmos. I. Universidade Federal do Paraná. II. Scarpin, Cassius Tadeu. III. Bonat, Wagner Hugo. IV. Título.

CDD: 519.536

Bibliotecário: Elias Barbosa da Silva CRB-9/1894

MINISTÉRIO DA EDUCAÇÃO
SETOR SETOR DE TECNOLOGIA
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO ENGENHARIA DE PRODUÇÃO - 40001016070P1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ENGENHARIA DE PRODUÇÃO da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **RICARDO RASMUSSEN PETTERLE** intitulada: **MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO**, após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 12 de Dezembro de 2018.

CASSIUS TADEU SCARPIN
Presidente da Banca Examinadora (UFPR)

MARCOS AUGUSTO MENDES MARQUES
Avaliador Interno (UFPR)

JOSE LUIZ PADILHA DA SILVA
Avaliador Externo (UFPR)

GUSTAVO VALENTIM LOCH
Avaliador Interno (UFPR)

Aos meus pais,
pelo apoio e incentivo.

AGRADECIMENTOS

Agradeço primeiramente à Deus, autor e princípio de tudo, que me proporcionou a sabedoria e a comunhão necessária para realização com êxito deste laborioso trabalho. Agradeço também à Nossa Senhora do Perpétuo Socorro, da qual sou um devoto eternamente agradecido.

Uma saudação especial a meus pais que, ao longo da vida, sempre estiveram ao meu lado, compartilhando minhas alegrias e apoiando-me em meus momentos difíceis.

Um agradecimento especial à minha amada esposa Roberta por toda sua paciência e compreensão, bem como aos demais familiares, que colaboraram de maneira positiva em meu processo de aquisição do conhecimento.

Agradeço ao meu orientador Professor Doutor Cassius Tadeu Scarpin por todo apoio e incentivo prestado durante a elaboração deste trabalho, além de toda sua ajuda para resolver questões burocráticas no PPGEP. Agradeço ao meu coorientador Professor Doutor Wagner Hugo Bonat, que desde minha graduação no curso de Estatística até os dias de hoje indicou-me o caminho do conhecimento, especialmente pelas ideias, conversas, discussões e paciência que teve comigo até a finalização deste. Agradeço também aos demais professores que estiveram comigo durante o processo formativo.

Agradeço a Professora Doutora Victória Zeghbi Cochenski Borba e à médica endocrinologista Thaísa Hoffmann Jonasson do Serviço de Endocrinologia e Metabologia do Hospital de Clínicas da Universidade Federal do Paraná pela colaboração com os dados do percentual de gordura corporal.

Não posso deixar de lembrar dos meus amigos e colegas, que sempre estiveram ao meu lado, e que mesmo sem perceber me deram forças para que este trabalho fosse concluído.

Finalmente, a todas as pessoas que, direta ou indiretamente, estiveram presentes na realização deste trabalho.

"A simplicidade é o último grau de sofisticação".
(Leonardo da Vinci)

RESUMO

Em diversas áreas de pesquisa é frequente a análise de dados com variáveis respostas limitadas ao intervalo unitário. Tais variáveis geralmente se apresentam na forma de taxas, proporções, indíces e porcentagens, sendo portanto limitadas ao intervalo $(0,1)$. Para o caso de múltiplas respostas é comum analisar cada variável resposta separadamente, o que não permite investigar possíveis correlações entre elas. Nesse sentido, o presente trabalho propõe um novo modelo de regressão para análise de variáveis respostas limitadas multivariada. O modelo é especificado usando apenas suposições de primeiro e segundo momentos. A abordagem usada para estimação dos parâmetros combina as funções de estimação quase-escore e Pearson para estimação dos parâmetros de regressão e dispersão, respectivamente. A principal vantagem da abordagem proposta é não precisar assumir uma distribuição de probabilidade multivariada para o vetor de variáveis respostas. O algoritmo de estimação é de fácil implementação, podendo ser resumido a um simples e eficiente algoritmo do tipo Newton-score. Além disso, o modelo proposto permite acomodar facilmente dados no intervalo $[0,1]$, incluindo excesso de zeros e uns. No decorrer do trabalho foram delineados três estudos de simulação. O primeiro foi conduzido para investigar o comportamento do algoritmo NORTA (*NORmal To Anything*) na simulação de variáveis aleatórias beta correlacionadas. O segundo visou explorar a flexibilidade dos estimadores para lidar com dados limitados em estudos longitudinais. E o terceiro foi delineado para checar propriedades dos estimadores como viés, consistência e taxa de cobertura em estudos com múltiplas respostas correlacionadas. O modelo foi motivado por dois conjuntos de dados que não são facilmente manipulados pelos métodos estatísticos convencionais. O primeiro se refere ao índice de qualidade da água de reservatórios de usinas hidrelétricas operadas pela COPEL no Estado do Paraná. E o segundo corresponde ao percentual de gordura corporal, que foi medido em cinco regiões do corpo e representam as variáveis respostas. Além disso, foram adaptadas técnicas de diagnóstico para o modelo proposto, tais como DFFITS, DFBETAS, distância de Cook e o gráfico de probabilidade meio-normal com envelope simulado, para detecção de pontos influentes e *outliers*. Portanto, as principais contribuições do modelo de regressão proposto nesta dissertação estão na análise de dados limitados em estudos longitudinais, além da análise de dados limitados em estudos com múltiplas respostas correlacionadas.

Palavras-chave: Múltiplas variáveis respostas limitadas. Dados correlacionados. Intervalo unitário. Dados longitudinais. Estudo de simulação. Algoritmo NORTA.

ABSTRACT

In several areas of research it is common to analyze data with response variables limited to the unit interval. These variables usually appear in the form of rates, proportions, index and percentages, being therefore limited to the interval $(0,1)$. When the response variable is multivariate, in general, each response variable is analyzed separately, which does not allow investigating possible correlations between them. Thus, we propose a multivariate regression model to deal with multiple continuous bounded data. The model is specified using only first and second moment assumptions and the method for estimation and inference combines the quasi-score and Pearson estimating functions for the estimation of the regression and dispersion parameters, respectively. The main advantage of the proposed approach is that it does not need to assume a multivariate probability distribution for the response vector. The fitting procedure is easily implemented using a simple and efficient Newton scoring algorithm. Furthermore, the proposed model can easily handle data in the unit interval, including exact zeros and ones. During the work, we conducted three simulation studies. The first one evaluated the behavior of the NORTA algorithm (NORmal To Anything) in the simulation of correlated beta random variables. The second aimed to explore the flexibility of estimators to deal with continuous bounded data in longitudinal studies. And the third was designed to check properties of the estimators, such as bias, consistency, and coverage rate in studies with multiple correlated response variables. The model was motivated by two data sets that are not easily manipulated by existing statistical methods. The first refers to the water quality index measured on power plant reservoirs operated by COPEL in the State of Paraná, Brazil. The second corresponds to the percentage of body fat, which was measured at five regions of the body and represent the response variables. We adapted diagnostic techniques for the proposed model, such as DFFITS, DFBETAS, Cook's distance and half-normal plot with simulated envelope, to check influential points and outliers. Therefore, the proposed model in this work allows the analysis of continuous bounded data in longitudinal studies, in addition to the analysis of continuous bounded data in studies with multiple correlated response variables.

Keywords: Multiple bounded response variables. Correlated data. Unit interval. Longitudinal data. Simulation study. NORTA algorithm.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em diversas áreas de pesquisa é comum investigar a relação entre uma variável de interesse com outras variáveis que compõem o estudo. Para tanto, faz-se uso da técnica estatística de modelos de regressão, uma vez que se pode estudar o relacionamento entre uma variável resposta (variável dependente) com possíveis variáveis explicativas (covariáveis) (MONTGOMERY; PECK; VINING, 2012). A aplicação desta técnica estatística é ampla, abrangendo diversas áreas do conhecimento como medicina, engenharias, agronomia, ciências sociais dentre outras. Nesse contexto, um dos principais modelos de regressão e sem dúvida um dos mais usados por usuários de estatística aplicada é o clássico modelo de regressão linear (Gaussiano). No entanto, para uso desse modelo alguns pressupostos devem ser atendidos, tais como erros independentes e identicamente distribuídos segundo a distribuição normal com média zero e variância constante (DRAPER; SMITH, 2014). Na prática, isso nem sempre acontece e a má especificação desse modelo pode gerar erros padrões inconsistentes, além de outros problemas que invalidam todo o processo de inferência (MYERS et al., 2010; MONTGOMERY; PECK; VINING, 2012). Apesar de amplamente utilizado, o modelo de regressão linear não é adequado para respostas binárias, politômicas, contagens ou limitadas.

Diante de tais limitações, Nelder e Wedderburn (1972) propuseram a classe de modelos lineares generalizados (GLM). Para ajuste dos GLMs é necessário que a distribuição da variável resposta pertença a família exponencial de distribuições. Dessa forma, os GLMs permitem a modelagem de respostas contínuas, contagens e binárias/binomial (NELDER; WEDDERBURN, 1972; MCCULLAGH; NELDER, 1989). Na sequência, Liang e Zeger (1986) propuseram o método de equações de estimação generalizadas (GEE) para analisar variáveis respostas correlacionadas não-Gaussianas. O método é especificado por um modelo de regressão marginal combinado com uma matriz de correlação de "trabalho" usada para modelar dependência. Desse modo, o método GEE estende a classe dos GLMs para análise de dados com medidas repetidas em estudos longitudinais (LIANG; ZEGER, 1986; ZEGER; LIANG; ALBERT, 1988).

Apesar da flexibilidade dos modelos acima mencionados, eles são pouco flexíveis para variáveis respostas cujo o suporte é limitado. Em geral, dados limitados se apresentam na forma de taxas, proporções, índices e porcentagens, sendo portanto limitados ao intervalo $(0,1)$. Exemplos de dados com tais características incluem: a proporção de itens defeituosos num lote, o índice de qualidade da água, a porcentagem de renda que uma família gasta com alimentação, o percentual de gordura corporal dentre outros.

Para analisar variáveis respostas limitadas, os modelos de regressão beta (FERRARI; CRIBARI-NETO, 2004) e simplex (KIESCHNICK; MCCULLOUGH, 2003) são escolhas usuais. Além destes modelos, outros foram propostos na literatura como os modelos de regressão gama unitário (MOUSA; EL-SHEIKH; ABDEL-FATTAH, 2016) e Johnson S_B (LEMONTE; BAZÁN, 2016). Recentemente, Bonat et al. (2018b) propuseram uma nova classe de modelos de regressão para análise de respostas limitadas. A abordagem proposta é baseada apenas em suposições de primeiro e segundo momentos com variância na forma $\phi\mu^p(1-\mu)^p$, onde μ é a média da variável resposta e ϕ e p são os parâmetros de dispersão e potência, respectivamente.

Embora os modelos supracitados possam ser usados em inúmeras aplicações, eles são limitados à análise de apenas uma variável resposta. Para o caso de duas ou mais variáveis respostas pode-se pensar em um modelo de regressão multivariado, o qual apresenta vantagens na análise dos dados e tem ganhado importância na literatura. A seguir, destacam-se as principais vantagens de um modelo de regressão multivariado:

1. Análise de múltiplas variáveis respostas conjuntamente;
2. Modelagem da matriz de correlação entre as respostas levando em conta o efeito das covariáveis no modelo;
3. Uso de testes de hipóteses e de comparações múltiplas multivariado como, por exemplo, a análise de variância multivariada (MANOVA).

Além do que foi descrito, espera-se que um modelo de regressão multivariado possa trazer mais informações na análise dos dados, além da flexibilidade para modelar dados com estruturas cada vez mais complexas. Nesse sentido, há um interesse crescente (BONAT et al., 2017), bem como a necessidade de implementação de novos métodos estatísticos para análise de dados com múltiplas respostas.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Propor um modelo de regressão para análise de variáveis respostas limitadas multivariada.

1.1.2 Objetivos específicos

1. Estudar o desempenho do algoritmo NORTA (*NORmal To Anything*) para simular variáveis aleatórias beta correlacionadas.
2. Especificar o modelo usando suposições de primeiro e segundo momentos.

3. Usar as funções de estimação quase-score e Pearson para estimar os parâmetros de regressão e dispersão, respectivamente.

4. Delinear estudos de simulação para explorar a flexibilidade do modelo para lidar com dados limitados em estudos longitudinais, além de checar propriedades dos estimadores em estudos com múltiplas respostas correlacionadas.

5. Adaptar técnicas de diagnóstico para o modelo proposto, como DFFITS, DFBETAS, distância de Cook e o gráfico de probabilidade meio-normal com envelope simulado.

6. Aplicar o modelo proposto em dois conjuntos de dados.

1.2 JUSTIFICATIVA

Variáveis respostas limitadas no intervalo $(0,1)$ apresentam particularidades na modelagem por regressão. Dados com essas características geralmente são assimétricos e se concentram perto das "bordas", isto é, perto do zero ou do um. Assim, é necessário um modelo de regressão adequado para tratar esse tipo de dado.

Quando a variável resposta é multivariada, em geral, cada variável resposta é analisada separadamente, o que não permite investigar possíveis correlações entre elas. Assim, a principal vantagem e contribuição do modelo de regressão proposto nesta dissertação está na análise de múltiplas respostas conjuntamente, além da análise de dados com medidas repetidas e dados agrupados em estudos longitudinais.

1.3 LIMITAÇÕES

Este trabalho se restringe a propor um novo modelo de regressão para análise de variáveis respostas limitadas multivariada. Para motivar o novo modelo, serão apresentadas aplicações em dois conjuntos de dados, que não são facilmente manipulados pelos métodos estatísticos existentes. Portanto, testes de hipóteses e de comparações múltiplas multivariados não serão desenvolvidos no decorrer deste trabalho.

1.4 ORGANIZAÇÃO DO TRABALHO

Esta dissertação contém seis capítulos incluindo esta introdução. O Capítulo 2 descreve os dois conjuntos de dados que serão usados como exemplos de aplicação no novo modelo. O Capítulo 3 apresenta a revisão bibliográfica que motivou este trabalho, introduz o modelo de regressão beta (univariado), apresenta o algoritmo NORTA (*NORmal To Anything*) usado nos estudos de simulação e discute brevemente as medidas de bondade de ajuste usadas no trabalho. O Capítulo 4 propõe o modelo de regressão

quase-beta multivariado, apresenta o método usado para estimação e inferência e adapta técnicas de diagnóstico. No Capítulo 5 são apresentados os resultados de três estudos de simulação, além da análise dos dados apresentados no Capítulo 2. Finalmente, o Capítulo 6 discute as principais contribuições desta dissertação, além de apresentar as conclusões seguidas por sugestões para futuros trabalhos.

2 CONJUNTOS DE DADOS

Este Capítulo descreve os dois conjuntos de dados que serão usados como exemplos de aplicação no novo modelo de regressão, proposto no Capítulo 4. O primeiro conjunto se refere ao índice de qualidade da água de reservatórios de usinas hidrelétricas operadas pela COPEL no Estado do Paraná. Já o segundo conjunto de dados corresponde ao percentual de gordura corporal de indivíduos avaliados no Hospital de Clínicas da Universidade Federal do Paraná.

2.1 CONJUNTO DE DADOS I: ÍNDICE DE QUALIDADE DA ÁGUA

O índice de qualidade da água (IQA) foi desenvolvido em 1970 nos Estados Unidos para avaliar a qualidade da água destinada ao abastecimento após seu tratamento. Em 1975, a CETESB (Companhia Ambiental do Estado de São Paulo) começou a utilizar este índice, sendo adotado mais tarde por outros Estados brasileiros como o principal indicador de qualidade da água. O IQA é calculado por meio de nove parâmetros físico-químicos e biológicos, considerados fundamentais para avaliação da qualidade da água (ABBASI; ABBASI, 2012). São eles: coliformes fecais, pH, demanda bioquímica de oxigênio, nitrogênio total, temperatura da água, fósforo total, turbidez, resíduo total e oxigênio dissolvido. Conforme o estado ou a condição de cada parâmetro, foram estabelecidas curvas de variação de qualidade da água (Anexo A) que mostram um conjunto de curvas médias com seus respectivos pesos (w) e valores de qualidade (q). Desse modo, o IQA é calculado pelo produtório ponderado dos valores de qualidade de cada parâmetro, resultando em uma nota entre zero e cem. Quanto maior for essa nota, melhor é a qualidade da água. Assim, define-se o cálculo do IQA pela Equação 2.1:

$$\text{IQA} = \prod_{i=1}^{9} q_i^{w_i}, \tag{2.1}$$

onde q_i (valor entre 0 e 100) corresponde a qualidade do i-ésimo parâmetro, obtido a partir do resultado das curvas médias e da análise de laboratório, e w_i (valor entre 0 e 1) se refere ao peso do i-ésimo parâmetro, de tal forma que $\sum_{i=1}^{9} w_i = 1$. Logo, o IQA apresenta fácil interpretação, principalmente, para um público leigo. Além disso, ele pode ser comparável com outras localidades. Por esse motivo, o IQA é usado por diversos institutos ambientais, como a Agência Nacional de Águas (ANA), o Instituto Ambiental do Paraná (IAP), dentre outros órgãos vinculados ao meio ambiente.

A Companhia Paranaense de Energia (COPEL) opera 16 usinas hidrelétricas no Estado do Paraná. A principal finalidade dessas usinas é geração de energia elétrica

por meio de rios e reservatórios de água. Além da geração de energia, a água dos reservatórios também é utilizada para outras finalidades, tais como: pesca, navegação, lazer, irrigação da agricultura e abastecimento das cidades. Para atender as especificações de operação dessas hidrelétricas, a COPEL monitora trimestralmente a qualidade da água a montante, a jusante e nos reservatórios dos rios represados. O principal objetivo desse monitoramento é detectar mudanças na qualidade da água, possivelmente atribuíveis à presença das barragens, além de preservar o meio ambiente e evitar a degradação das águas. A Figura 1 ilustra uma das usinas operadas pela COPEL e seus respectivos pontos de coleta da água.

FIGURA 1 – USINA HIDRELÉTRICA DA COPEL E PONTOS DE COLETA DA ÁGUA

FONTE: Copel (2018).

O estudo foi conduzido em 2004, e avaliou 12 medidas (3 locais $\times$ 4 trimestres) em cada uma das 16 usinas, resultando num total de 190 observações com apenas dois dados faltantes (*missing data*). O principal objetivo da análise dos dados foi investigar o relacionamento do IQA com os locais (montante, reservatório e jusante) controlado pelo efeito dos trimestres e das usinas. Dessa forma, tem-se características de um estudo longitudinal e de dados agrupados. A primeira característica está relacionada com os trimestres, enquanto a segunda está associada com os locais. Geralmente, um estudo longitudinal permite analisar mudanças na variável resposta ao longo do tempo, além de investigar o efeito de covariáveis (DIGGLE et al., 2002). Além disso, pode-se analisar uma possível correlação intraunidades amostrais, estudando a estrutura da matriz de covariância (DIGGLE et al., 2002; FITZMAURICE; LAIRD; WARE, 2011).

No problema em questão, a principal abordagem, e sem dúvida a mais comum entre usuários de estatística aplicada é a análise de variância (ANOVA). Porém, deve-se atentar para seus pressupostos, tais como observações independentes, homogeneidade entre grupos e resíduos independentes provenientes de uma distribuição Gaussiana com média zero e variância constante (MONTGOMERY; PECK; VINING, 2012). Um dos principais problemas em usar esse método está na forma como os trimestres e locais são avaliados, isto é, eles são tratados como observações independentes na amostra. Outro problema está na natureza da variável resposta, que é limitada ao intervalo unitário e portanto representa particularidades na análise dos dados.

Possivelmente as observações deste estudo são correlacionadas. Dessa forma, precisa-se de um método estatístico adequado para investigar tal questão. Assim, o modelo de regressão proposto no Capítulo 4 será usado para análise dos dados, uma vez que esse modelo permite considerar uma estrutura de covariância para análise de dados longitudinais e outra para dados agrupados.

A Tabela 1 apresenta a análise descritiva para o IQA, com valores expressos em média e desvio padrão $(\bar{x} \pm s)$, conforme os trimestres e locais. Desta tabela, observa-se que o reservatório foi o local que apresentou os maiores valores do IQA. Além disso, nos trimestres 2 e 3 o IQA apresentou-se superior aos demais trimestres.

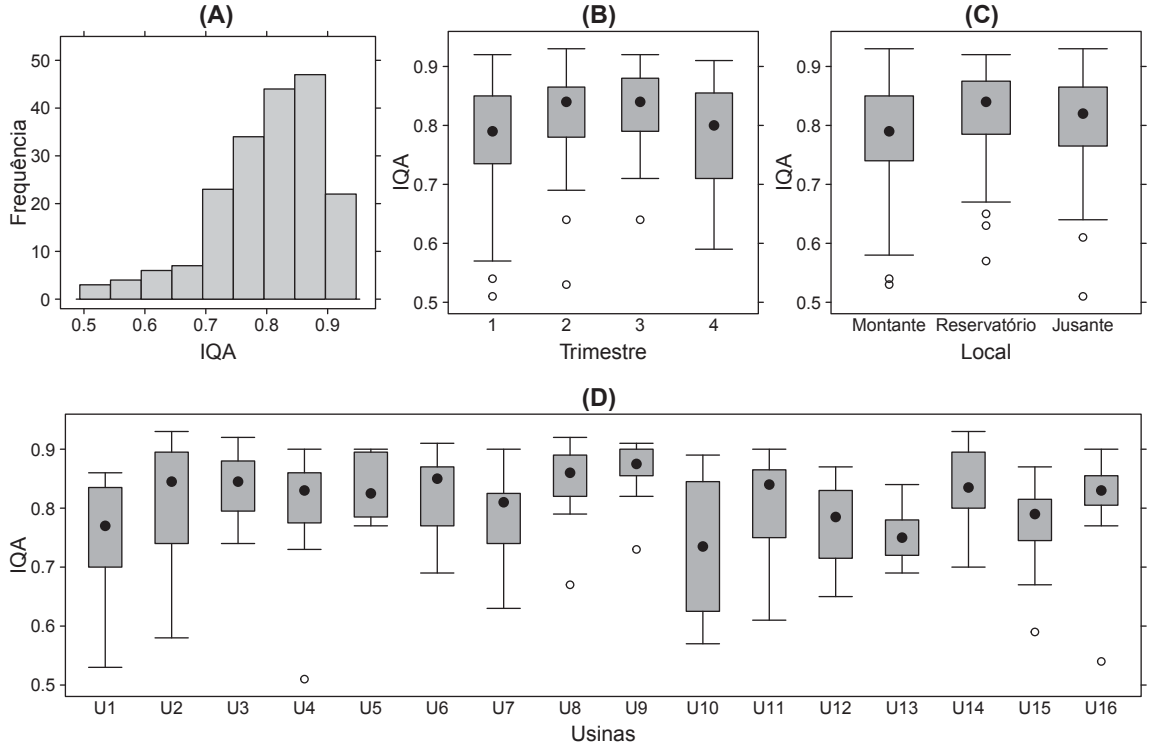
TABELA 1 – ANÁLISE DESCRITIVA PARA O IQA POR TRIMESTRE E LOCAL

Trimestre	Local		
	Montante	Reservatório	Jusante
1	$0,75 \pm 0,11$	$0,80 \pm 0,10$	$0,78 \pm 0,10$
2	$0,79 \pm 0,10$	$0,83 \pm 0,06$	$0,83 \pm 0,07$
3	$0,81 \pm 0,07$	$0,85 \pm 0,05$	$0,83 \pm 0,06$
4	$0,76 \pm 0,10$	$0,81 \pm 0,08$	$0,79 \pm 0,09$

FONTE: O autor (2018).

A Figura 2 complementa as informações da Tabela 1, apresentando um histograma e diagramas boxplot para o IQA de acordo com as covariáveis do estudo. A Figura 2 (A) sugere distribuição assimétrica à esquerda para o IQA, enquanto a Figura 2 (B) indica valores maiores para o IQA durante os trimestres 2 e 3. Já a Figura 2 (C) indica que o IQA foi maior no reservatório do que nos outros locais, confirmando portanto o que foi observado na Tabela 1.

FIGURA 2 – HISTOGRAMA (A) E BOXPLOTS PARA O ÍNDICE DE QUALIDADE DA ÁGUA (IQA) POR TRIMESTRE (B), LOCAL (C) E USINAS (D)



FONTE: O autor (2018).

Por fim, os resultados apresentados na Figura 2 (D) mostram que o IQA não é homogêneo entre as usinas, com um destaque maior para as usinas 1, 2 e 10. É importante ressaltar que os resultados apresentados na Tabela 1 e Figura 2 se referem apenas a análise descritiva e exploratória dos dados, onde são criadas hipóteses que serão confirmadas somente após ajuste do modelo de regressão proposto no Capítulo 4. No Apêndice A são apresentados gráficos boxplots para o IQA separado por trimestre e local em função das usinas.

2.2 CONJUNTO DE DADOS II: PERCENTUAL DE GORDURA CORPORAL

Este estudo foi realizado no Serviço de Endocrinologia e Metabologia (SEMPR) do Hospital de Clínicas (HC) da Universidade Federal do Paraná (UFPR). Trata-se de um estudo observacional, transversal com voluntários saudáveis recrutados após divulgação da pesquisa no HC-UFPR, consultórios, salas de aula e entre familiares. Após informações sobre a pesquisa os participantes assinaram o termo de consentimento livre e esclarecido (TCLE), aprovado pelo Comitê de Ética em Pesquisa do mesmo hospital. Foram incluídos homens e mulheres saudáveis, entre 18 e 90 anos, sem uso de drogas ou derivados hormonais, seja para reposição ou suplementação, com índice de massa corporal (IMC) entre 18,5 e 29,9 Kg/m^2, sem qualquer incapacidade física e que

andassem sem ajuda de órteses ou próteses. Foram excluídos os indivíduos portadores de doenças crônicas e medicamentos ou drogas lícitas ou ilícitas que, sabidamente, afetem a composição corporal, como diabetes insulino-dependente, corticoesteróides, hormônio tireoidiano em doses supressivas e aqueles com baixo peso, condizente com IMC menor que 18,5 Kg/m^2 ou obesos, com IMC de no mínimo 30 Kg/m^2. Logo, a amostra utilizada neste estudo contém 298 observações.

Todos os participantes realizaram no mesmo dia medidas antropométricas (peso e altura) e responderam ao questionário sobre dados sociodemográficos, seguida do exame de densitometria de corpo total (Aparelho *Lunar Prodigy Advance* PA+302284) para análise das massas gorda, magra e óssea do corpo total. O exame foi avaliado segundo a recomendação da *International Society for Clinical Densitometry* (ISCD) (PETAK et al., 2013; KENDLER et al., 2013).

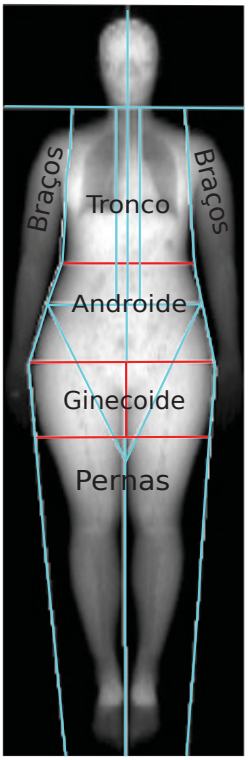
Todos os participantes responderam ao IPAQ (*international physical activity questionnaire*), validado em português (MATSUDO et al., 2001), que é um instrumento para estimar o nível de atividade física praticada habitualmente. Foi utilizado o IPAQ-curto, composto por oito questões sobre realização, frequência e duração de atividades físicas moderadas, vigorosas ou de caminhada. Levando-se em conta, para as respostas, a última semana ou uma semana habitual de exercícios físicos. O IPAQ foi respondido na forma de autoadministração para a maioria dos voluntários ou como entrevista individual, aplicada pelo investigador ou avaliador treinado, nos casos em que houve dificuldade de entendimento. Os voluntários foram, então, divididos em três grupos, conforme o nível de atividade física realizada (NAHAS, 2001; SONATI, 2012): sedentários são aqueles que não realizam nenhuma atividade física por pelo menos 10 minutos contínuos durante a semana; insuficientemente ativos, realizam no mínimo 10 minutos contínuos de atividade física, pelo menos 5 dias na semana ou 150 minutos por semana, porém de maneira insuficiente para serem classificados como ativos. Os ativos são os indivíduos que realizam no mínimo 20 minutos de atividade física vigorosa por sessão, pelo menos 3 vezes na semana ou atividades moderadas, ou caminhada de 30 minutos por sessão, pelo menos 5 vezes na semana ou qualquer atividade somada por 5 dias da semana ou mais, com duração total de 150 minutos por semana (SILVA et al., 2007).

O estudo é composto por quatro covariáveis: sexo (F-feminino ou M-masculino), idade (anos), IMC (Kg/m^2) e IPAQ (S-sedentário, IA-insuficientemente ativo ou A-ativo). Nesta dissertação, analisou-se o percentual de gordura corporal que foi medido em cinco regiões do corpo e representam as variáveis respostas. A Figura 3 apresenta uma ilustração dessas regiões, que estão divididas em braços, pernas, tronco, andróide e ginecóide. Cabe ressaltar, que a imagem apresentada na Figura 3 é o resultado do exame de densitometria de corpo total, realizado no aparelho *Lunar Prodigy Advance* PA+302284.

Por meio de tal exame, foi possível mensurar o percentual de gordura corporal de cada indíviduo, além de outras medidas relativas a composição corporal que não fazem parte deste estudo. Desse modo, os percentuais de massas gorda, magra e óssea foram analisados separadamente em um outro estudo, fazendo-se uso do modelo de regressão beta. Para detalhes, ver Petterle et al. (2018).

A amostra contém 150 mulheres e 148 homens. Os indivíduos têm média de idade de 46 anos com desvio-padrão de 19,88. O IMC foi estimado em $24,72 \pm 3,15$ Kg/m^2. Segundo o questionário IPAQ, que avalia o nível de atividade física, 76 indivíduos foram classificados em insuficientemente ativos, 60 em sedentários e 162 em ativos.

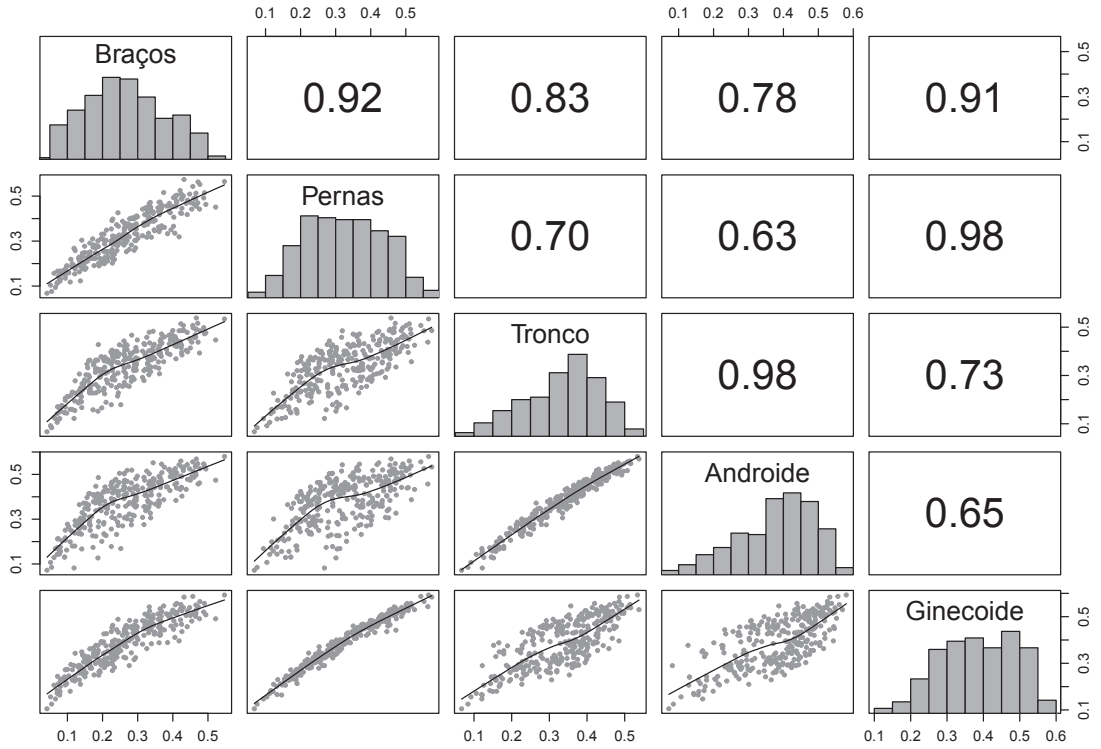
FIGURA 3 – RESULTADO DO EXAME DE DENSITOMETRIA DE CORPO TOTAL DIVIDIDO POR REGIÕES DO CORPO



FONTE: O autor (2018).

A Figura 4 apresenta diagramas de dispersão com curvas de suavização estimadas pelo método *loess* (CLEVELAND, 1979), além de mostrar as correlações entre os percentuais de gordura nas regiões dos braços, pernas, tronco, androide e ginecoide. De acordo com os resultados apresentados na Figura 4, todos os coeficientes de correlação estimados são positivos podendo-se observar correlações mais fortes entre os percentuais de gordura nas regiões dos braços e pernas $(\hat{\rho} = 0,92)$, tronco e androide $(\hat{\rho} = 0,98)$, braços e ginecoide $(\hat{\rho} = 0,91)$ e entre pernas e ginecoide $(\hat{\rho} = 0,98)$.

FIGURA 4 – DIAGRAMAS DE DISPERSÃO E CORRELAÇÕES ENTRE OS PERCENTUAIS DE GORDURA NA REGIÃO DOS BRAÇOS, PERNAS, TRONCO, ANDROIDE E GINECOIDE

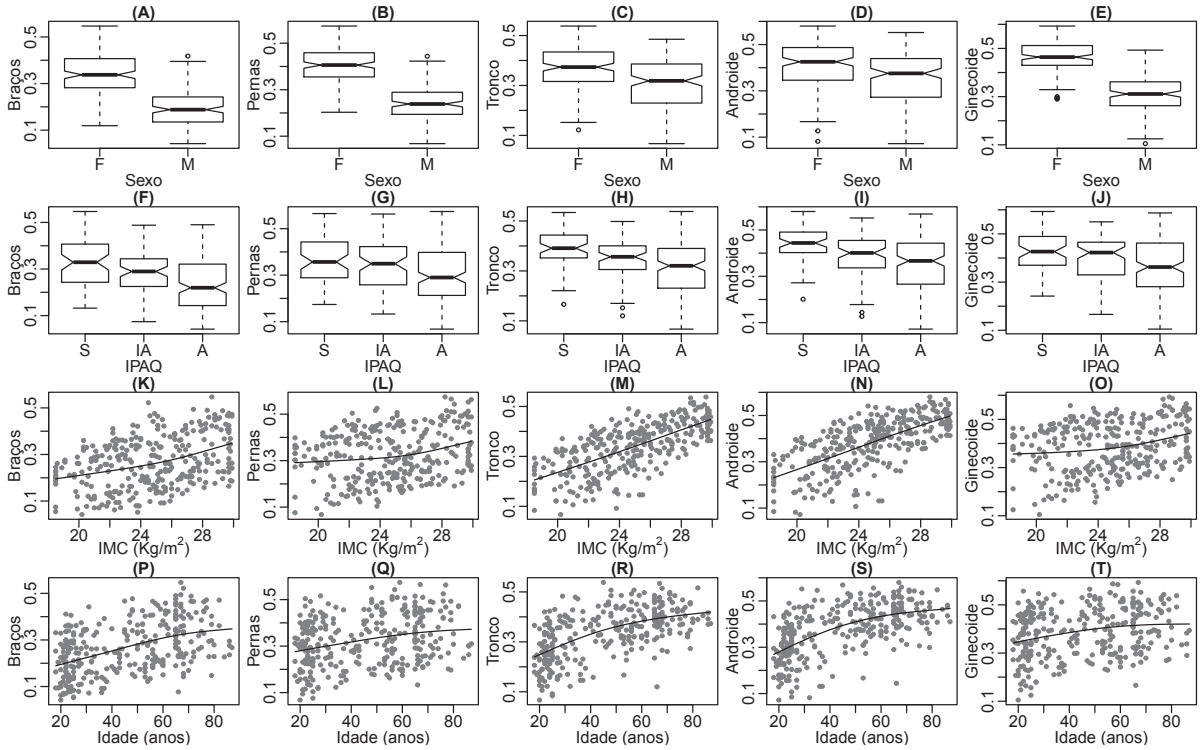


FONTE: O autor (2018).

Por outro lado, correlações mais fracas podem ser observadas entre os percentuais de gordura nas regiões das pernas e androide ($\hat{\rho} = 0,63$) e entre androide e ginecoide ($\hat{\rho} = 0,65$). É importante ressaltar que essas correlações foram estimadas pelo coeficiente de correlação de Spearman ($\hat{\rho}$) e não levam em conta o efeito marginal das covariáveis presentes no estudo. Adicionalmente, a Figura 4 mostra a distribuição de cada variável resposta por meio de um histograma, indicando distribuições simétricas para a maioria delas. No entanto, distribuições assimétricas à esquerda podem ser vistas para os percentuais de gordura nas regiões do tronco e androide, respectivamente.

A Figura 5 apresenta gráficos boxplots e diagramas de dispersão para cada variável resposta em função das covariáveis do estudo. Com base nos resultados apresentados na Figura 5 (A-E) tem-se um indicativo de que os percentuais de gordura corporal diferem entre homens e mulheres, com um destaque maior para as regiões dos braços, pernas e ginecoide. A Figura 5 (F-J) indica que os percentuais de gordura corporal diminuem à medida que os indivíduos praticam atividade física. A Figura 5 (K-O) mostra correlações positivas para o IMC dos indivíduos, destacando-se os percentuais de gordura nas regiões do tronco e andróide que apresentam tendência linear e, possivelmente, correlações mais fortes. Finalmente, a Figura 5 (P-T) indica que todos os percentuais de gordural corporal aumentam gradativamente com avançar da idade.

FIGURA 5 – GRÁFICOS BOXPLOTS PARA SEXO (A-E) E IPAQ (F-J). DIAGRAMAS DE DISPERSÃO PARA OS PERCENTUAIS DE GORDURA NA REGIÃO DOS BRAÇOS, PERNAS, TRONCO, ANDROIDE E GINECOIDE POR IMC (K-O) E IDADE (P-T)



FONTE: O autor (2018).

É importante destacar que os resultados apresentados na Figura 5 se referem apenas a análise descritiva dos dados, onde são apontadas as principais e possíveis covariáveis associadas com as variáveis respostas. Por meio do modelo de regressão multivariado proposto no Capítulo 4, espera-se avaliar conjuntamente esses resultados, além de investigar possíveis correlações entre as variáveis respostas, dada a presença das covariáveis no modelo. De acordo com Verbeke et al. (2014), muitas perguntas de pesquisa podem ser respondidas modelando as inúmeras variáveis respostas do estudo separadamente, porém, algumas questões só podem ser respondidas em uma análise conjunta de todas elas.

Logo, o principal objetivo da análise dos dados é investigar a relação entre os percentuais de gordura corporal (braços, pernas, tronco, androide e ginecoide) com sexo, nível de atividade física (IPAQ), IMC e idade dos indivíduos. O objetivo secundário da análise é estimar a matriz de correlação entre as respostas, dada a presença das covariáveis no modelo.

3 FUNDAMENTAÇÃO TEÓRICA

Este Capítulo apresenta a fundamentação teórica que será usada nesta dissertação. A seção 3.1 apresenta um breve resumo dos principais trabalhos relacionados ao assunto. A distribuição de probabilidade beta e suas propriedades encontram-se na seção 3.2. A seção 3.3 apresenta o algoritmo NORTA, que será usado para simular variáveis aleatórias beta correlacionadas. A seção 3.4 introduz o modelo de regressão beta (univariado). Por fim, a seção 3.5 apresenta brevemente as medidas de bondade de ajuste usadas na comparação entre os modelos.

3.1 REVISÃO DA LITERATURA

Variáveis respostas limitadas são comuns em diversas áreas de pesquisa, como em medicina, ciências sociais, engenharias dentre outras. Paolino (2001) foi um dos primeiros autores a propor uma distribuição de probabilidade para modelar variáveis aleatórias limitadas ao intervalo $(0,1)$. Na sequência, Kieschnick e McCullough (2003) apresentaram diversos modelos de regressão para análise de respostas limitadas, dentre eles o modelo linear gaussiano com censura, os modelos de regressão beta e simplex e um modelo semi-paramétrico estimado por quase-verossimilhança. Ferrari e Cribari-Neto (2004) apresentaram o popular modelo de regressão beta a partir da reparametrização na densidade beta, que permite modelar diretamente sua resposta média em função de covariáveis. Extensões do modelo beta são descritas em Smithson e Verkuilen (2006), Simas, Barreto-Souza e Rocha (2010) e Grün, Kosmidis e Zeileis (2012). Além disso, alguns autores usaram o modelo de regressão simplex como forma alternativa ao modelo de regressão beta (LÓPEZ, 2013). Miyashiro (2008) comparou os modelos de regressão beta e simplex com aplicações em dois conjuntos de dados reais. Bonat, Ribeiro Jr e Zeviani (2012) especificaram uma classe geral de modelos a partir da escolha da distribuição de probabilidade, função de ligação e/ou transformação para a variável resposta. Nesta classe de modelos, os autores incluíram a densidade beta (FERRARI; CRIBARI-NETO, 2004), simplex (BARNDORFF-NIELSEN; JØRGENSEN, 1991), Kumaraswamy (LEMONTE; BARRETO-SOUZA; CORDEIRO, 2013), gaussiana e gaussiana transformada.

Recentemente, Liu e Eugenio (2018) revisaram e compararam aspectos metodológicos e computacionais de inferência bayesiana e por máxima verossimilhança nos modelos de regressão beta e beta inflacionada, incluindo modelos com/sem efeitos aleatórios. Ainda, para o caso de uma variável resposta, outros modelos foram propostos na literatura para análise de respostas limitadas. Exemplos incluem o modelo de

regressão gama unitário desenvolvido por Mousa, El-Sheikh e Abdel-Fattah (2016) e o modelo de regressão Johnson S_B apresentado por Lemonte e Bazán (2016). Além destes, Mitnik e Baek (2013) apresentaram um modelo de regressão baseado na distribuição Kumaraswamy, que permite modelar diretamente a mediana da variável resposta em função de covariáveis. Adicionalmente, Bonat et al. (2018b) propuseram uma nova classe de modelos de regressão para análise de variáveis respostas limitadas. Os modelos são baseados em suposições de segundo momentos, isto é, média e variância, onde μ é a média da variável resposta e $\phi\mu^p(1-\mu)^p$ é a variância. Nessa notação, ϕ é o parâmetro de dispersão e p é um parâmetro de potência, que foi introduzido para dar mais flexibilidade na modelagem da relação entre média e variância (BONAT et al., 2018b).

No contexto de dados limitados em estudos longitudinais, Song e Tan (2000), Song, Qiu e Tan (2004) e Qiu, Song e Tan (2008) apresentaram uma classe de modelos marginais baseados na distribuição simplex, com dispersão fixa e variável. Verkuilen e Smithson (2012) usaram o modelo de regressão beta com efeitos aleatórios na análise de experimentos de psicologia cognitiva, enquanto Hunger, Döring e Holle (2012) mostraram aplicações na área médica. Já Bonat, RIBEIRO JR e Zeviani (2015) usaram a mesma abordagem para discutir métodos de inferência por máxima verossimilhança com aplicações a dados reais. Sob o paradigma de inferência Bayesiana, o modelo beta com efeitos mistos foi discutido em Figueroa-Zúñiga, Arellano-Valle e Ferrari (2013). Masarotto, Varin et al. (2012) propuseram uma classe de modelos marginais baseados em cópulas gaussianas, incluindo o modelo marginal beta. Sob a estrutura de séries temporais alguns modelos foram propostos (MCKENZIE, 1985; GRUNWALD; RAFTERY; GUTTORP, 1993; ROCHA; CRIBARI-NETO, 2008; SILVA; MIGON; CORREIA, 2011; BAYER; BAYER; PUMI, 2017). Recentemente, Zhao, Lian e Bandyopadhyay (2018) apresentaram um modelo aditivo parcialmente linear para análise de dados limitados correlacionados, enquanto Zheng, Qin e Tu (2017) usaram uma abordagem similar para analisar dados de qualidade de vida de pacientes com câncer.

Embora os modelos supracitados possam ser usados em inúmeras aplicações, eles são limitados à análise de apenas uma variável resposta. Em certos casos, o pesquisador pode estar interessado em analisar mais de uma variável resposta simultaneamente. Além disso, pode existir interesse em investigar possíveis correlações entre as respostas, dada a presença de covariáveis no modelo. Acredita-se que um modelo de regressão multivariado possa agregar mais informações na análise dos dados, produzindo estimativas dos parâmetros mais precisas e confiáveis. Nesse sentido, alguns autores propuseram modelos de regressão para modelagem conjunta de variáveis respostas limitadas ao intervalo unitário.

O modelo de regressão beta bivariado com modelagem conjunta de média e

dispersão foi proposto por Cepeda-Cuervo, Achcar e Lopera (2014). Os autores usaram a função copula Farlie–Gumbel–Morgenstern (FGM) para construção da distribuição beta bivariada, em que os parâmetros do modelo são estimados conjuntamente via inferência bayesiana. Souza e Moura (2016) propuseram duas classes de modelos de regressão multivariados com resposta beta. O primeiro usa cópulas para a modelagem marginal das variáveis respostas, enquanto o segundo modelo considera a inclusão de efeitos fixos e aleatórios, resultando num modelo hierárquico multivariado com efeitos aleatórios correlacionados. Já o modelo de regressão Dirichlet é uma outra abordagem quando se tem mais de uma variável resposta. Esse modelo foi usado inicialmente por Hijazi e Jernigan (2009) e explorado mais tarde por Murteira e Ramalho (2016) com aplicações na área econômica, onde os autores apresentam uma reparametrização do modelo, além de estudos de simulação.

Cabe ressaltar, que o modelo de regressão Dirichlet é comumente usado na análise de dados composicionais, em que um vetor de respostas $Y_1, Y_2, \ldots, Y_p$, denominadas composições, representam p-frações de algum "inteiro" e satisfazem a restrição de soma igual a um, isto é, $Y_1 + Y_2 + \ldots + Y_p = 1$ e $Y_1 \geq 0, Y_2 \geq 0, \ldots, Y_p \geq 0$ (AITCHISON, 1986). Devido a tal restrição, essas variáveis não são mutuamente exclusivas, ou seja, quando o valor de uma variável se altera, obrigatoriamente os valores das outras variáveis também são alterados. Logo, o modelo de regressão Dirichlet é diferente do modelo de regressão proposto nesta dissertação e, portanto, não se aplica ao conjunto de dados apresentado na seção 2.2.

3.2 DISTRIBUIÇÃO DE PROBABILIDADE BETA

A distribuição de probabilidade beta é um modelo probabilístico para variáveis aleatórias contínuas e restritas ao intervalo $(0,1)$. Essa distribuição de probabilidade possui dois parâmetros, ambos positivos, que controlam o formato da distribuição. Considere Y uma variável aleatória beta, cuja função densidade de probabilidade é dada por:

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1},\ p>0,\ q>0, \tag{3.1}$$

onde $0 < y < 1$ e $\Gamma(\cdot)$ denota a função gama. Sua média e variância são dadas, respectivamente, por:

$$\mathrm{E}(Y) = \frac{p}{p+q} \quad \text{e} \quad \mathrm{Var}(Y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

Em modelagem por regressão é de interesse relacionar a média da variável resposta com uma ou mais covariáveis. Assim, Ferrari e Cribari-Neto (2004) propuseram uma reparametrização da densidade beta, na qual pode-se modelar diretamente sua resposta média em função de covariáveis, $\mu = p/(p+q)$. Além disso, incluiu-se o

parâmetro $\phi = p + q$, sendo este um parâmetro de dispersão. Reescrevendo, tem-se: $p = \mu\phi$ e $q = (1-\mu)\phi$. Consequentemente, sua densidade fica da seguinte forma:

$$f(y,\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{3.2}$$

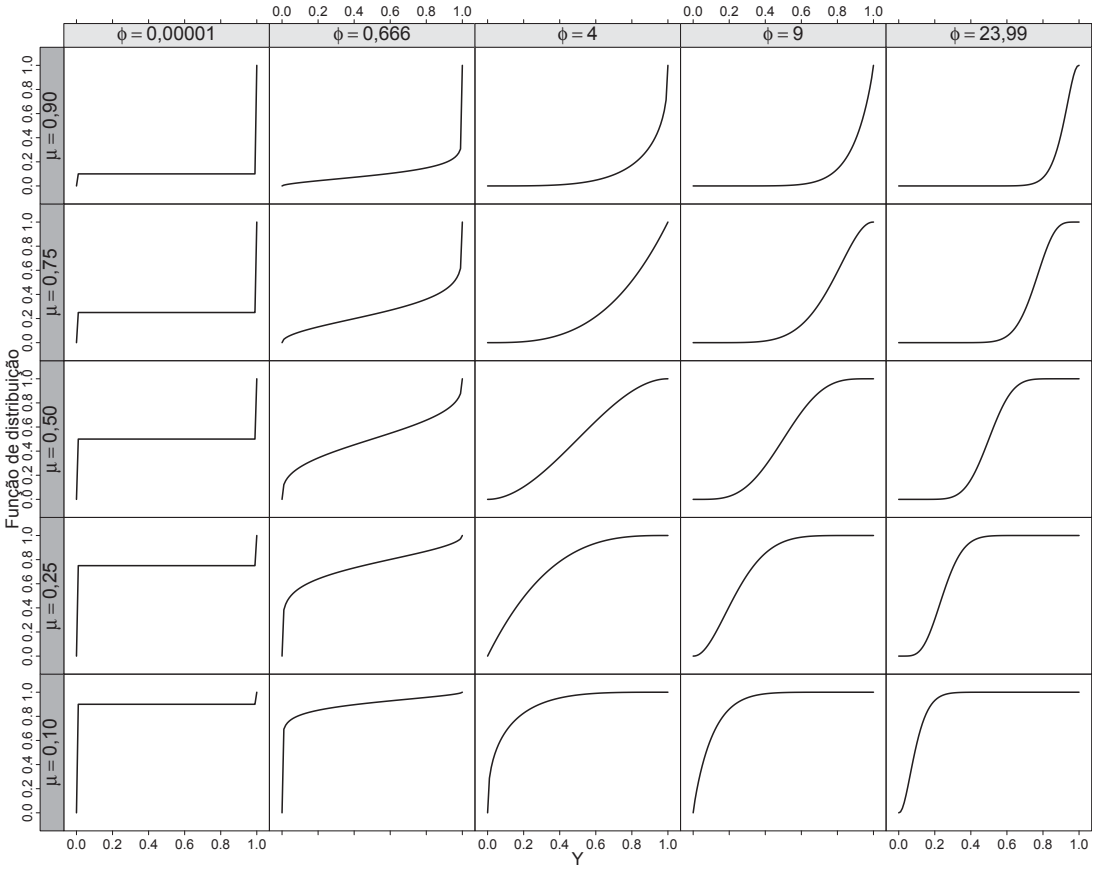
onde $0 < \mu, y < 1$ e $\phi > 0$. Logo, a média e a variância da densidade beta na sua forma reparametrizada são dadas, respectivamente, por:

$$\mathrm{E}(Y) = \mu \quad \text{e} \quad \mathrm{Var}(Y) = \frac{\mu(1-\mu)}{1+\phi}.$$

Note que, a $\mathrm{Var}(Y) \to \mu(1-\mu)$ quando $\phi \to 0$. Dessa forma, tem-se como caso particular a relação entre média e variância da distribuição Bernoulli.

A Figura 6 apresenta a função de distribuição beta para diferentes valores de μ combinados com $\phi = (0{,}00001;\ 0{,}666;\ 4;\ 9;\ 23{,}99)$. Tais valores foram definidos quando $\mu = 0{,}5$ e $\mathrm{Var}(Y) = 0{,}25;\ 0{,}15;\ 0{,}05;\ 0{,}025;\ 0{,}01$, respectivamente.

FIGURA 6 – FUNÇÃO DE DISTRIBUIÇÃO BETA PARA DIFERENTES VALORES DE μ COMBINADOS COM $\phi = (0{,}00001;\ 0{,}666;\ 4;\ 9;\ 23{,}99)$



FONTE: O autor (2018).

Os resultados apresentados na Figura 6 mostram que para valores mais altos de ϕ a função de distribuição beta fica mais concentrada em torno da média. Além disso, na medida em que $\phi \to \infty$ a função de distribuição beta tende para o caso Gaussiano.

3.3 ALGORITMO NORTA

O algoritmo NORTA (CARIO; NELSON, 1997) é um dos métodos mais populares para simulação de vetores aleatórios correlacionados. O método funciona como um processo em dois passos. Primeiro é gerado um vetor aleatório normal multivariado $\mathbf{Z}$ de dimensão $p \times 1$. Logo então, esse vetor é transformado num vetor uniforme multivariado $\mathbf{U}$, que é novamente transformado no vetor $\mathbf{Y}$ que tem distribuição NORTA (*NORmal To Anything*), onde cada elemento do vetor tem distribuição marginal arbitrária desejada. Logo, sua representação é dada por:

$$\mathbf{Y} = \left[F_{Y_1}^{-1}(\Phi[Z_1]), F_{Y_2}^{-1}(\Phi[Z_2]), \ldots, F_{Y_p}^{-1}(\Phi[Z_p])\right]^{\top}, \tag{3.3}$$

onde $\Phi[\cdot]$ é a função de distribuição acumulada da distribuição normal padrão aplicada a cada elemento do vetor $\mathbf{Z}$ e $F_{Y_l}^{-1}(u) \equiv \inf\{y : F_{Y_l}(y) \geq u\}$ é a inversa da função de distribuição acumulada.

A matriz de correlação de $\mathbf{Z}$ determina diretamente a matriz de correlação de $\mathbf{Y}$, desde que

$$\rho_Y(l,l') = \mathrm{Corr}(Y_l, Y_{l'}) = \mathrm{Corr}\big(F_{Y_l}^{-1}(\Phi[Z_l]), F_{Y_{l'}}^{-1}(\Phi[Z_{l'}])\big),$$

para todo $l \neq l'$. A correlação é definida por:

$$\mathrm{Corr}(Y_l, Y_{l'}) = \frac{\mathrm{E}(Y_l, Y_{l'}) - \mathrm{E}(Y_l)\mathrm{E}(Y_{l'})}{\sqrt{\mathrm{Var}(Y_l)\mathrm{Var}(Y_{l'})}}, \tag{3.4}$$

onde as quantidades marginais $\mathrm{E}(Y_l), \mathrm{E}(Y_{l'}), \mathrm{Var}(Y_l)$ e $\mathrm{Var}(Y_{l'})$ são definidas por F_{Y_l} e $F_{Y_{l'}}$. Vale destacar que $(Z_l, Z_{l'})$ tem distribuição normal padrão bivariada com correlação $\mathrm{Corr}(z_l, z_{l'}) = \rho_Z(l,l')$ onde a quantidade $\mathrm{E}(Y_l, Y_{l'})$ em (3.4) é calculada por:

$$\begin{aligned} \mathrm{E}(Y_l, Y_{l'}) &= \mathrm{E}\Big(F_{Y_l}^{-1}(\Phi[Z_l]) F_{Y_{l'}}^{-1}(\Phi[Z_{l'}])\Big) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{Y_l}^{-1}(\Phi[Z_l]) F_{Y_{l'}}^{-1}(\Phi[Z_{l'}]) \varphi_{\rho_Z(l,l')}(z_l, z_{l'}) dz_l dz_{l'}, \end{aligned} \tag{3.5}$$

onde $\varphi_{\rho_Z(l,l')}$ denota a função densidade de probabilidade de uma distribuição normal padrão bivariada com correlação dada por $\rho_Z(l,l')$.

Cabe ressaltar, que nem sempre a integral (3.5) vai ter solução, devido à relação média/variância da distribuição beta. Tal restrição, impacta diretamente no espaço paramétrico da correlação (3.4) que também depende da especificação das médias marginais.

3.3.1 Resumo do algoritmo NORTA

- **Entrada**: Distribuições marginais desejadas $F_{Y_l}(y), l = 1,2,\ldots,p$ e matriz de correlação $\mathbf{\Sigma}_Y = \mathrm{Corr}(Y_l, Y_{l'})$.

- **Saída**: Vetor aleatório Y com as marginais desejadas $F_{Y_l}(y)$ e sua respectiva matriz de correlação $\mathbf{\Sigma}_Y$.

  - **Passo 1**: Gere um vetor aleatório $\mathbf{Z}$ com distribuição normal p-variada com vetor de média zero e matriz de correlação $\mathbf{\Sigma}_Z$, isto é, $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_Z)$.
  - **Passo 2**: Avalie $\mathbf{Y} = \left[Y_1, Y_2, \ldots, Y_p\right]^\top$ via $Y_l = F_{Y_l}^{-1}(\Phi[Z_l]), l = 1,2,\ldots,p$.

3.3.2 Algoritmo NORTA no software R

O algoritmo NORTA pode ser facilmente implementado em diversos *softwares* estatísticos e linguagens de programação. Neste trabalho, fez-se uso do *software* estatístico `R` (R DEVELOPMENT CORE TEAM, 2018) para gerar variáveis aleatórias beta correlacionadas. Para exemplificar sua aplicação, considere os códigos mostrados na Figura 7. Tais códigos visam ilustar os passos apresentados na subseção 3.3.1, para gerar um conjunto de dados com 1000 observações a partir de uma distribuição beta bivariada com $\mu = 0{,}5$ e $\phi = 9$ para cada distribuição marginal e correlação fixa em $\rho = 0{,}75$.

É importante observar, que nesta subseção os códigos `R` encontram-se em figuras para melhor visualização. Além disso, comentários feitos no código estão à direita do símbolo #.

FIGURA 7 – CÓDIGOS EM LINGUAGEM R PARA GERAÇÃO DE VARIÁVEIS ALEATÓRIAS BETA CORRELACIONADAS

```
R = 1000  # tamanho da amostra
mu = 0.5  # parâmetro de média
phi = 9  # parâmetro de dispersão
cor_matrix <- matrix(c(1.0,0.75,0.75,1.0),2,2) # matriz de correlação
require(MASS)  # carrega o pacote com a função mvrnorm()
Z <- mvrnorm(n = R, mu = c(0,0), Sigma = cor_matrix)  # passo 1
Y <- qbeta(pnorm(Z), shape1 = mu*phi, shape2 = (1 - mu)*phi) # passo 2
```

FONTE: O autor (2018).

Alternativamente, o pacote `NORTARA` (SU, 2014) do *software* estatístico `R` fornece toda implementação computacional do algoritmo NORTA. Por meio da função `genNORTARA()` é possível gerar o vetor aleatório $\mathbf{Y}$ com as marginais desejadas.

O uso do pacote facilita quando os parâmetros das distribuições marginais são diferentes. Além disso, na geração do vetor aleatório NORTA, as distribuições marginais não precisam necessariamente serem iguais. É possível simular dados com um misto de distribuições discretas e contínuas. Outra utilidade do pacote é a função

`valid_input_cormat()`, que retorna os valores mínimo (ρ_L) e máximo (ρ_U) que a matriz de correlação pode assumir em função das distribuições marginais. A aplicação desta função será discutida mais adiante (subseção 5.1.1), onde um estudo de simulação foi conduzido para investigar o comportamento do algoritmo NORTA na simulação de variáveis aleatórias beta correlacionadas.

A reprodução do exemplo anterior (Figura 7), por meio do pacote `NORTARA`, é apresentada nos códigos da Figura 8.

FIGURA 8 – CÓDIGOS EM LINGUAGEM R PARA GERAÇÃO DE VARIÁVEIS ALEATÓRIAS BETA CORRELACIONADAS USANDO O PACOTE NORTARA

```
require(NORTARA)  # carrega o pacote
invcdfnames <- c("qbeta","qbeta")  # define as marginais beta
cor_matrix <- matrix(c(1.0,0.75,0.75,1.0),2,2) # matriz de correlação
mu = 0.5  # parâmetro de média
phi = 9  # parâmetro de dispersão
paramslists <- list(m1 = list(shape1 =  mu*phi, shape2 = (1 - mu)*phi),
                    m2 = list(shape1 = mu*phi, shape2 = (1 - mu)*phi))
Y <- genNORTARA(1000,cor_matrix,invcdfnames,paramslists)  # gera os dados
```

FONTE: O autor (2018).

Na sequência, a Figura 9 mostra as 6 primeiras linhas do vetor `Y`, composto pelas marginais beta correlacionadas.

FIGURA 9 – RESULTADO GERADO PELO ALGORITMO NORTA

```
head(Y)
          [,1]      [,2]
[1,] 0.4946017 0.5187232
[2,] 0.6256042 0.4959388
[3,] 0.2841912 0.2495754
[4,] 0.6202264 0.4859596
[5,] 0.3681033 0.3272861
[6,] 0.1995514 0.2199893
```

FONTE: O autor (2018).

Note que, cada coluna de `Y` define as densidades marginais beta Y_1 e Y_2, respectivamente.

3.4 MODELO DE REGRESSÃO BETA

O modelo de regressão beta foi introduzido por Ferrari e Cribari-Neto (2004) com objetivo de modelar variáveis respostas contínuas pertencentes ao intervalo $(0,1)$. Esse modelo foi estruturado a partir da reparametrização da densidade beta que permite modelar diretamente sua resposta média, bem como seu parâmetro de dispersão em função de covariáveis. Para detalhes sobre a densidade beta na sua forma reparametrizada, ver seção 3.2.

Seja $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes, em que cada variável Y_i $(i = 1,2,\ldots,n)$ segue a função de densidade beta (3.2), isto é, $Y_i \sim \mathcal{B}(\mu_i, \phi)$, de tal forma que o modelo de regressão beta é definido por $f(y;\mu,\phi)$ e pela seguinte relação funcional:

$$g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta},$$

onde $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^\top$ é o vetor $(k \times 1)$ das covariáveis (conhecidas), $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^\top$ é o vetor $(k \times 1)$ dos coeficientes de regressão (desconhecidos) e $g(\cdot)$ é uma função de ligação. As principais funções de ligação usadas na análise de dados limitados são: *logit*, *probit*, *cloglog* e *cauchit*. Além dessas, outras funções de ligação podem ser adaptadas para análise de dados limitados como, por exemplo, as funções de ligação Aranda-Ordaz, Weibull, Prentice e Stukel. É importante observar, que tais ligações são consideradas paramétricas, uma vez que dependem de parâmetros extras. Para detalhes, ver Santos (2013) e Petterle et al. (2017).

A função de ligação *logit* é uma das mais populares entre usuários de estatística aplicada, pois permite interpretar os coeficientes de regressão em termos de razão de chances (RC). Para isso, basta exponenciar os coeficientes β_j, ou seja, $\text{RC} = \exp\{\beta_j\}$ para $j = 1,\ldots,k$. Além disso, pode-se construir intervalos de confiança conforme Equação 3.6:

$$\exp\left\{\beta_j \pm z_{1-\alpha/2}\text{EP}(\beta_j)\right\}, \tag{3.6}$$

onde $z_{1-\alpha/2}$ é um ponto obtido a partir da distribuição normal padrão e $\text{EP}(\beta_j)$ é o erro-padrão associado a estimativa de β_j.

3.5 MEDIDAS DE BONDADE DE AJUSTE

A seleção e comparação de modelos é uma questão importante em quase toda análise de dados. O modelo de regressão proposto nesta dissertação não apresenta uma função de verossimilhança explícita. Desse modo, para fazer comparações entre modelos foram usadas medidas de bondade de ajuste (*goodnees-of-fit*) propostas por Bonat (2018). Tais medidas se referem aos pseudo critérios de informação de *Akaike* (`pAIC`) e

Bayesiano (pBIC), assim como o valor maximizado do logaritmo da função de pseudo verossimilhança (plogLik).

Quanto menor forem os valores dos pseudo critérios de informação de *Akaike* (pAIC) e Bayesiano (pBIC) mais parcimoniosos são os modelos. Por outro lado, quanto maior for o valor do logaritmo da função de pseudo verossimilhança (plogLik) melhor é o ajuste do modelo. Para maiores detalhes, ver Bonat (2018).

4 MODELO DE REGRESSÃO MULTIVARIADO

Este Capítulo apresenta o novo modelo de regressão usado para análise de variáveis respostas limitadas multivariada, o qual será chamado por modelo de regressão quase-beta multivariado. A seção 4.1 apresenta a estrutura do modelo, enquanto a seção 4.2 apresenta o método proposto para estimação dos parâmetros de regressão e dispersão. Por fim, a seção 4.3 adapta técnicas de diagnóstico para o modelo proposto.

4.1 MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO

Considere $\mathbf{Y}_i = (\mathrm{Y}_{1i}, \ldots, \mathrm{Y}_{Ri})^\top$ um vetor de variáveis respostas $R \times 1$ e seja $\boldsymbol{\mu}_i = (\mu_{1i}, \ldots, \mu_{Ri})^\top$ seu correspondente vetor de médias $R \times 1$, para $r = 1, \ldots, R$ variáveis respostas e $i = 1, \ldots, n$ indivíduos. Nesta notação, a média da r-ésima variável resposta para o i-ésimo indivíduo é dada por

$$\mu_{ri} = g_r^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}_r),$$

onde $\mathbf{x}_i$ e $\boldsymbol{\beta}_r$ são $J_r \times 1$ vetores de covariáveis conhecidas e desconhecidos parâmetros de regressão e g_r é uma função de ligação. Logo, o modelo de regressão quase-beta multivariado é definido por

$$\begin{aligned} \mathrm{E}(\mathbf{Y}_i) &= \boldsymbol{\mu}_i \\ \mathrm{Var}(\mathbf{Y}_i) &= \boldsymbol{\Sigma}_i = \mathrm{V}(\boldsymbol{\mu}_i)^{\frac{1}{2}} \boldsymbol{\Omega}(\boldsymbol{\lambda}) \mathrm{V}(\boldsymbol{\mu}_i)^{\frac{1}{2}} \end{aligned} \tag{4.1}$$

onde $\boldsymbol{\Sigma}_i$ é uma matriz $R \times R$ e $\mathrm{V}(\boldsymbol{\mu}_i)$ denota uma matriz diagonal cujas entradas principais são dadas por $\vartheta(\mu_{ri}) = \mu_{ri}(1 - \mu_{ri})$. Como este modelo é baseado apenas em suposições de segundo momentos ele pode ser expresso, alternativamente, da seguinte forma

$$\begin{pmatrix} \mathrm{Y}_{1i} \\ \vdots \\ \mathrm{Y}_{Ri} \end{pmatrix} \sim \bullet \left[\begin{pmatrix} \mu_{1i} \\ \vdots \\ \mu_{Ri} \end{pmatrix}; \boldsymbol{\Sigma}_i \right], \; i = 1, \ldots, n.$$

Note que, o modelo não assume uma distribuição de probabilidade multivariada para o vetor de variáveis respostas, sendo que a notação $\bullet$ substitui tal suposição. Para análise de dados limitados com múltiplas respostas, a matriz $\boldsymbol{\Sigma}_i$ tem a seguinte representação:

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sqrt{\mu_{1i}(1-\mu_{1i})} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\mu_{Ri}(1-\mu_{Ri})} \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \ldots & \rho_{1R}\sigma_1\sigma_R \\ \vdots & \ddots & \vdots \\ \rho_{1R}\sigma_1\sigma_R & \ldots & \sigma_R^2 \end{bmatrix} \begin{bmatrix} \sqrt{\mu_{1i}(1-\mu_{1i})} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\mu_{Ri}(1-\mu_{Ri})} \end{bmatrix},$$

onde ρ_{1R}, σ_R^2 e σ_R, para $r = 1, \ldots, R$ são os parâmetros de correlação, variância (dispersão) e covariância, respectivamente.

É importante destacar, que a matriz $\boldsymbol{\Omega}(\boldsymbol{\lambda})$ descreve a parte da covariância que não depende da estrutura de média. A ideia é semelhante a abordagem por equações de estimação generalizadas proposta por Liang e Zeger (1986) e Zeger, Liang e Albert (1988), que usam uma matriz de correlação de "trabalho" para modelar dependência. Essa estrutura de dependência, geralmente é usada na análise de dados longitudinais, dados com medidas repetidas e dados agrupados. Dessa forma, a matriz $\boldsymbol{\Omega}(\boldsymbol{\lambda})$ pode ser customizada para análise de dados com tais características. Nesse caso, por simplicidade, substituiu-se a notação e adaptou-se a estrutura de covariância, trocando $\boldsymbol{\lambda}$ por $\boldsymbol{\tau}$. Assim, tal estrutura é definida por:

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \Lambda_0 + \ldots + \tau_Q \Lambda_Q, \tag{4.2}$$

onde Λ_q com $q = 0, \ldots, Q$ são matrizes conhecidas que refletem a estrutura de interesse, e $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_Q)$ é um vetor de parâmetros. Bonat et al. (2017), seguindo os argumentos de Demidenko (2013), mostraram que abordagens populares para lidar com dados longitudinais autocorrelacionados, como os modelos permutável, médias móveis e autorregressiva de primeira ordem são modelos lineares de covariância, ou seja, têm a forma de (4.2). Cabe ressaltar que em modelos lineares de covariância, como apresentado na Equação 4.2, o principal desafio é definir restrições para o espaço paramétrico de cada componente do vetor $\boldsymbol{\tau}$. Assim, a única restrição imposta para a matriz $\boldsymbol{\Sigma}_i$ é que ela seja positiva definida. Logo, a especificação do preditor linear matricial (4.2) para diferentes estruturas de covariância será discutida no estudo de simulação apresentado na subseção 5.1.2, bem como na análise dos dados na subseção 5.2.1.

Assim, o modelo de regressão proposto nesta dissertação segue o estilo de quase-verossimilhança apresentado por Wedderburn (1974), que combina a função de variância da distribuição binomial com as tradicionais funções de ligação para dados binários como, por exemplo, as ligações *logit*, *probit*, *cloglog* e *cauchit*, além de uma estrutura de covariância, especificada pela combinação linear de matrizes conhecidas.

4.2 ESTIMAÇÃO E INFERÊNCIA

Nesta seção, apresenta-se uma visão geral do método usado para estimação dos parâmetros do modelo de regressão quase-beta multivariado. A abordagem proposta é baseada em funções de estimação. Para maiores detalhes sobre o método e principais resultados, ver Jørgensen e Knudsen (2004) e Bonat e Jørgensen (2016). Assim, para estimar os parâmetros do modelo, a abordagem proposta combina as funções de estimação quase-score e Pearson para a estimação dos parâmetros de regressão e dispersão, respectivamente.

Seja $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\lambda}^\top)^\top$ um vetor composto por dois conjuntos de parâmetros, em que $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_R^\top)$ e $\boldsymbol{\lambda} = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_R^2, \rho_1, \ldots, \rho_{R(R-1)/2})^\top$ são vetores $J \times 1$ e $Q \times 1$ de parâmetros associados aos coeficientes de regressão e dispersão, respectivamente. A função quase-escore para $\boldsymbol{\beta}$ é definida por:

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} \mathbf{D}_i^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i),$$

onde $\mathbf{D} = \nabla_{\boldsymbol{\beta}} \boldsymbol{\mu}_i$ é uma matriz $R \times J$ e $\nabla_{\boldsymbol{\beta}}$ denota o operador gradiente. A matriz de sensibilidade $J \times J$ de $\psi_{\boldsymbol{\beta}}$ é dada por

$$\mathrm{S}_{\boldsymbol{\beta}} = -\sum_{i=1}^{n} \mathbf{D}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i, \tag{4.3}$$

onde a soma é elemento-a-elemento. De maneira análoga, a matriz de variabilidade $J \times J$ para $\psi_{\boldsymbol{\beta}}$ é obtida por

$$\mathrm{V}_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{D}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i.$$

A função de estimação de Pearson foi usada para estimar os parâmetros de dispersão e de acordo com Jørgensen e Knudsen (2004), Bonat e Jørgensen (2016), Bonat et al. (2018a) ela tem a seguinte forma:

$$\psi_{\lambda_q}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \mathrm{tr} \left\{ W_{i\lambda_q} \left[\Delta_i^\top \Delta_i - \boldsymbol{\Sigma}_i \right] \right\}, \; q = 1, \ldots, Q,$$

onde o operador tr denota o traço da matriz, $\Delta_i = \mathbf{Y}_i - \boldsymbol{\mu}_i$ e $W_{i\lambda_q} = -\partial \boldsymbol{\Sigma}_i^{-1} / \partial \boldsymbol{\lambda}_q$. Detalhes sobre o cálculo da matriz $W_{i\lambda_q}$ podem ser vistos em Bonat e Jørgensen (2016) Seção 3.1.

A entrada (q, q') da matriz $Q \times Q$ de sensitividade para $\psi_{\boldsymbol{\lambda}}$ é dada por

$$\mathrm{S}_{\lambda_{qq'}} = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\lambda}_q} \psi_{\lambda_{q'}}(\boldsymbol{\lambda}, \boldsymbol{\beta}) \right) = -\sum_{i=1}^{n} \mathrm{tr}\Big(W_{i\lambda_q} \boldsymbol{\Sigma}_i W_{i\lambda_{q'}} \boldsymbol{\Sigma}_i \Big). \tag{4.4}$$

A matriz de sensitividade cruzada para $\boldsymbol{\beta}$ e $\boldsymbol{\lambda}$ é dada por

$$\mathrm{S}_{\beta_j \lambda_q} = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\lambda}_q} \psi_{\beta_j}(\boldsymbol{\beta}, \boldsymbol{\lambda}) \right) = \mathbf{0} \tag{4.5}$$

e

$$\mathrm{S}_{\lambda_q \beta_j} = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta_j}} \psi_{\lambda_q}(\boldsymbol{\lambda}, \boldsymbol{\beta}) \right) = -\sum_{i=1}^{n} \mathrm{tr}\Big(W_{i\lambda_q} \boldsymbol{\Sigma}_i W_{i\beta_j} \boldsymbol{\Sigma}_i \Big), \tag{4.6}$$

em que $W_{i\beta_j} = -\partial \boldsymbol{\Sigma}_i^{-1} / \partial \boldsymbol{\beta}_j$. A matriz de sensitividade conjunta para o vetor $\boldsymbol{\theta}$ fica representada por

$$\mathrm{S}_{\boldsymbol{\theta}} = \begin{pmatrix} \mathrm{S}_{\boldsymbol{\beta}} & \mathbf{0} \\ \mathrm{S}_{\lambda\boldsymbol{\beta}} & \mathrm{S}_{\boldsymbol{\lambda}} \end{pmatrix},$$

cujas entradas são definidas pelas Equações (4.3)-(4.6).

A variância assintótica dos estimadores baseados em funções de estimação denotado por $\hat{\boldsymbol{\theta}}$ é obtido pela inversa da matriz de informação Godambe, cuja forma geral para o vetor de parâmetros $\boldsymbol{\theta}$ é $\mathrm{J}_{\boldsymbol{\theta}}^{-1} = \mathrm{S}_{\boldsymbol{\theta}}^{-1}\mathrm{V}_{\boldsymbol{\theta}}\mathrm{S}_{\boldsymbol{\theta}}^{-\top}$, onde $^{-\top}$ denota a transposta inversa, isto é, $\mathrm{S}_{\boldsymbol{\theta}}^{-\top} = (\mathrm{S}_{\boldsymbol{\theta}}^{-1})^{\top}$. A matriz de variabilidade para $\boldsymbol{\theta}$ tem a forma

$$\mathrm{V}_{\boldsymbol{\theta}} = \begin{pmatrix} \mathrm{V}_{\boldsymbol{\beta}} & \mathrm{V}_{\boldsymbol{\beta\lambda}} \\ \mathrm{V}_{\boldsymbol{\lambda\beta}} & \mathrm{V}_{\boldsymbol{\lambda}} \end{pmatrix} \tag{4.7}$$

onde $\mathrm{V}_{\boldsymbol{\lambda\beta}} = \mathrm{V}_{\boldsymbol{\beta\lambda}}^{\top}$ e $\mathrm{V}_{\boldsymbol{\lambda}}$ dependem do terceiro e quarto momentos de $\mathbf{Y}_i$, respectivamente.

As entradas para matriz de variabilidade empírica são dadas por:

$$\tilde{\mathrm{V}}_{\lambda_{qq'}} = \sum_{i=1}^{n} \psi_{\lambda_j}(\boldsymbol{\lambda},\boldsymbol{\beta})_i \psi_{\lambda_{q'}}(\boldsymbol{\lambda},\boldsymbol{\beta})_i \quad \text{e} \quad \tilde{\mathrm{V}}_{\lambda_q\beta_{q'}} = \sum_{i=1}^{n} \psi_{\lambda_{q'}}(\boldsymbol{\lambda},\boldsymbol{\beta})_i \psi_{\beta_{q'}}(\boldsymbol{\lambda},\boldsymbol{\beta})_i.$$

Denote $\hat{\boldsymbol{\theta}}$ o estimador função de estimação de $\boldsymbol{\theta}$. De acordo com Godambe e Thompson (1978), Jørgensen e Knudsen (2004) a distribuição assintótica de $\hat{\boldsymbol{\theta}}$ é dada por

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \mathrm{J}_{\boldsymbol{\theta}}^{-1}),$$

onde $\mathrm{J}_{\boldsymbol{\theta}}^{-1} = \mathrm{S}_{\boldsymbol{\theta}}^{-1}\mathrm{V}_{\boldsymbol{\theta}}\mathrm{S}_{\boldsymbol{\theta}}^{-\top}$ é o inverso da matriz de informação Godambe.

O Algoritmo Chaser foi proposto por Jørgensen e Knudsen (2004) para resolver o sistema de equações $\psi_{\boldsymbol{\beta}} = \mathbf{0}$ e $\psi_{\boldsymbol{\lambda}} = \mathbf{0}$ e sua definição é dada por:

$$\begin{aligned} \boldsymbol{\beta}^{(i+1)} &= \boldsymbol{\beta}^{(i)} - \mathrm{S}_{\boldsymbol{\beta}}^{-1}\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)}) \\ \boldsymbol{\lambda}^{(i+1)} &= \boldsymbol{\lambda}^{(i)} - \alpha\mathrm{S}_{\boldsymbol{\lambda}}^{-1}\psi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}). \end{aligned} \tag{4.8}$$

A principal característica do algoritmo chaser é a propriedade de insensitividade (4.5), que permite separar $\boldsymbol{\beta}$ e $\boldsymbol{\lambda}$ em duas equações para serem atualizadas em cada passo. Além disso, o parâmetro α foi introduzido como um *tuning* constante para controlar o tamanho do passo, que é definido pelo usuário e varia entre 0 e 1. É importante observar que esse algoritmo é um caso especial do algoritmo flexível apresentado por Bonat e Jørgensen (2016), no contexto de modelos lineares generalizados multivariados com estrutura de covariância (McGLMs), e foi adaptado para o modelo proposto nesta dissertação. Desse modo, toda a implementação computacional do modelo foi feita no *software* estatístico `R` (R DEVELOPMENT CORE TEAM, 2018) e está disponível no pacote `mcglm` (BONAT, 2016) por meio da função `mcglm()`.

4.3 TÉCNICAS DE DIAGNÓSTICO

Uma etapa fundamental no ajuste de modelos de regressão é a análise de diagnóstico. O principal objetivo desta etapa é validar os resultados produzidos pelo

modelo, além de identificar as possíveis causas de desajuste. Assim, é possível avaliar a qualidade gerada pelo ajuste do modelo e, ainda, verificar se os seus pressupostos estão sendo atendidos.

De acordo com Montgomery, Peck e Vining (2012), os resíduos são definidos pela diferença entre os valores observados e preditos (ajustados) pelo modelo. Na literatura estatística foram desenvolvidos diversos tipos de resíduos, tais como os resíduos ordinários, padronizados, studentizados dentre outros. Além disso, foram desenvolvidas técnicas de diagnóstico afim de detectar possíveis pontos influentes e de alavanca, além de outros recursos diagnóstico como o gráfico de probabilidade meio-normal com envelope simulado (ATKINSON, 1985).

De maneira geral, a análise de resíduos e as técnicas de diagnóstico consistem na construção e avaliação de gráficos para um determinado tipo de resíduo ou medida diagnóstico. Testes de hipóteses também podem ser considerados para complementação desta etapa da análise, porém, a realização de tais testes demandam procedimentos mais formais. Talvez a principal desvantagem na análise de resíduos, está na familiaridade que o analista de dados deve ter com os diferentes tipos de resíduos e medidas diagnóstico, de tal forma que ele seja capaz interpretá-los adequadamente, além de ser capaz de identificar as possíveis causas de má especificação do modelo.

Nesta dissertação, adaptou-se os resíduos de Pearson, a distância de Cook (COOK, 1977) além de medidas como DFFITS, DFBETAS (BELSLEY; KUH; WELSCH, 1980) e o gráfico de probabilidade meio-normal com envelope simulado (ATKINSON, 1985) para o modelo proposto neste Capítulo (seção 4.1). Para detalhes sobre tais medidas, no contexto de análise de dados longitudinais, ver Venezuela, Botter e Sandoval (2007).

Para simplificar a notação, suponha $R = 1$ e p o número de coeficientes de regressão. Um ponto fundamental para o cálculo das medidas supracitadas é a matriz de projeção bloco diagonal $\mathbf{H} = \text{Bdiag}(\mathbf{H}_1, \ldots, \mathbf{H}_n)$, também conhecida por matriz chapéu, nos quais seus elementos são definidos por:

$$\mathbf{H}_i = \boldsymbol{\Sigma}_i^{1/2}\mathbf{X}_i(\mathbf{X}^\top\boldsymbol{\Sigma}\mathbf{X})^{-1}\mathbf{X}_i^\top\boldsymbol{\Sigma}_i^{1/2}, \text{ para } i = 1, \ldots, n, \tag{4.9}$$

onde $\mathbf{X}_i$ é a matriz de delineamento, $\boldsymbol{\Sigma} = \text{Bdiag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n)$ é uma matriz bloco diagonal com dimensão $n \times n$, onde $\boldsymbol{\Sigma}_i$ é a matriz de covariância do modelo de regressão quase-beta multivariado (Equação 4.1) e $\boldsymbol{\Sigma}_i^{1/2}$ é sua correspondente matriz raiz quadrada, obtida pelo Teorema da Decomposição Espectral.

A matriz $\mathbf{H}$ possui duas propriedades que a tornam interessante, ela é simétrica e idempotente. Isso significa que sua matriz transposta ($\mathbf{H}^\top = \mathbf{H}$), além do produto dela por ela mesma ($\mathbf{HH} = \mathbf{H}$) resultam na própria matriz $\mathbf{H}$. Dessa forma, tem-se que $\text{posto}(\mathbf{H}) = \text{tr}(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = p$. Assim, os elementos h_{ii} são fundamentais para a

construção de medidas de diagnóstico, tais como DFFITS, DFBETAS e distância de Cook. Note que, os elementos h_{ii} são obtidos a partir da diagonal principal da matriz de projeção e, em geral, são usados para avaliar a infuência do valor observado de y_i sobre seu correpondente valor estimado $\hat{y}_i$. Em outras palavras, o elemento h_{ii} pode ser interpretado como a variação em $\hat{y}_i$ quando acrescenta-se um valor infinitamente pequeno em y_i, ou seja, $h_{ii} = \partial\hat{y}_i/\partial y_i$. Desse modo, espera-se que os elementos h_{ii} estejam próximos de p/n, pois acredita-se que todos os pontos exerçam a mesma influência sobre os valores ajustados. Nesse sentido, deve-se examinar com cuidado aqueles pontos em que $h_{ii} > 2p/n$, sendo estes considerados pontos de alavanca.

Os resíduos de Pearson também são usados para avaliar o ajuste de um modelo de regressão, além de serem úteis para identificar padrões sistemáticos de variação. Nesta dissertação, eles foram adaptados para o modelo proposto, sendo obtidos por:

$$(r_P)_i = \frac{(Y_i - \hat{Y}_i)}{\sqrt{\text{diag}(\hat{\boldsymbol{\Sigma}}_i)}}, \tag{4.10}$$

onde $\hat{Y}_i$ são os valores preditos pelo modelo e $\hat{\boldsymbol{\Sigma}}_i$ denota a matriz de covariância do modelo de regressão quase-beta multivariado. A avaliação dos resíduos de Pearson é feita por meio de um gráfico, onde os valores obtidos na Equação 4.10 são plotados versus os valores de $\hat{Y}_i$. Em geral, nessa avaliação, valores entre -2 e 2 indicam que o ajuste do modelo está adequado aos dados.

Denote $(r_{SD})_i$ os resíduos padronizados dados por:

$$(r_{SD})_i = \frac{\mathbf{e}_i^\top \hat{\boldsymbol{\Sigma}}_i^{-1/2}(Y_i - \hat{\mu}_i)}{\sqrt{1-h_{ii}}},\ i = 1,\ldots,n, \tag{4.11}$$

onde $\mathbf{e}_i^\top$ é um vetor $n \times 1$ com o valor 1 na posição referente à observação Y_i e zero nas demais posições. Tal resíduo é fundamental para a construção de técnicas de diagnóstico, como a distância de Cook e o gráfico de probabilidade meio-normal com envelope simulado. De acordo com Cook (1977) e Venezuela, Botter e Sandoval (2007) a distância de Cook é definida por:

$$\text{CD}_i = (r_{SD})_i^2 \frac{h_{ii}}{p(1-h_{ii})}, \tag{4.12}$$

onde $(r_{SD})_i$ são os resíduos padronizados, h_{ii} são os elementos da diagonal principal da matriz $\mathbf{H}$ e p é o número de coeficientes de regressão estimados pelo modelo. A distância de Cook é visualizada graficamente, fazendo-se CD_i versus o índice de observações i.

A medida DFFITS foi proposta inicialmente por Belsley, Kuh e Welsch (1980) para avaliar da infuência da observação i sobre seu próprio valor estimado no ajuste de um modelo de regressão linear. Dessa forma, tal medida é usada para avaliar o quanto y_i é afetado pela presença ou ausência da i-ésima observação. Para construção

da medida DFFITS, no contexto de modelos de regressão estimados por funções de estimação, considere $(r_{ST})_i$ os resíduos studentizados dados por:

$$(r_{ST})_i = \frac{(Y_i - \hat{\mu}_i)}{\hat{\sigma}_i\sqrt{(1-h_{ii})}}, \tag{4.13}$$

onde $\hat{\sigma}_i = \sqrt{\text{diag}(\hat{\boldsymbol{\Sigma}}_i)}$. Desse modo, a medida DFFITS é dada por:

$$\text{DFFITS}_i = (r_{ST})_i\sqrt{\frac{h_{ii}}{1-h_{ii}}}. \tag{4.14}$$

Em geral, a avaliação é feita por meio de um gráfico dos DFFITS_i versus o número de observações i, nos quais valores fora do intervalo $\pm 2\sqrt{\frac{p}{n}}$ devem ser investigados.

A medida DFBETAS é uma técnica de diagnóstico comumente usada para avaliar a alteração do coeficiente de regressão estimado, em unidades padronizadas, ocasionada pela remoção da i-ésima observação. Tal medida é dada por

$$\text{DFBETAS}_i = \frac{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(Y_i - \hat{\mu}_i)}{\hat{\sigma}_i\sqrt{\text{diag}(\mathbf{X}^\top\mathbf{X})^{-1}}}\frac{1}{(1-h_{ii})}. \tag{4.15}$$

A avaliação desta medida também é feita visualmente por meio de um gráfico dos DFBETAS_i versus o índice de observações i, uma vez que valores fora do intervalo $\pm 2/\sqrt{n}$ merecem atenção.

Outra técnica de diagnóstico usada com frequência é o gráfico de probabilidade meio-normal com envelope simulado (*half-normal plot with simulated envelope*). De acordo com Neter et al. (1996), essa técnica é útil para identificar *outliers* e avaliar o ajuste produzido pelo modelo, mesmo quando a distribuição do resíduos não seja conhecida. A construção de um gráfico de probabilidade meio-normal é feita com base na ordenação do i-ésimo $(i = 1,\ldots,n)$ valor absoluto do resíduo padronizado (Equação 4.11) versus o valor esperado da estatística de ordem meio normal, na qual pode ser aproximada por:

$$\Phi^{-1}\left(\frac{i+n-\frac{1}{8}}{2n+\frac{1}{2}}\right), \tag{4.16}$$

onde $\Phi(\cdot)$ corresponde a função de distribuição acumulada da distribuição normal padrão. A seguir, encontram-se os passos usados para a construção do gráfico de probabilidade meio-normal com envelope simulado, no qual foi adaptado para o modelo de regressão proposto nesta dissertação.

1. Ajuste o modelo de regressão quase-beta multivariado;

2. Defina a r-ésima variável resposta que se deseja fazer o gráfico;

3. Calcule os resíduos padronizados conforme a Equação 4.11 e ordene seus valores absolutos;

4. Simule uma variável resposta com distribuição beta e adote a seguinte parametrização: $p = \hat{\mu}\hat{\phi}$ e $q = (1 - \hat{\mu})\hat{\phi}$. Note que, $\hat{\mu}$ corresponde aos valores preditos pelo modelo de regressão quase-beta multivariado e o parâmetro de dispersão é obtido por $\hat{\phi} = (1 - \hat{\tau})/\hat{\tau}$.

5. Ajuste o modelo de regressão quase-beta multivariado L vezes com a resposta simulada no passo anterior. Então, calcule os resíduos padronizados e ordene seus valores absolutos. Conforme indicam Tan, Qu e Kutner (1997) e Venezuela, Botter e Sandoval (2007), em geral, adota-se $L = 25$ repetições.

6. Calcule o mínimo, a mediana e o máximo dos menores valores absolutos dos resíduos padronizados de todas as simulações, isto é, para $l = 1, \ldots, L$.

7. Repita o passo anterior n vezes, calculando-se do menor valor absoluto dos resíduos das simulações até o maior valor. Após a finalização deste passo, são obtidos três vetores com dimensão $n \times 1$ contendo os valores absolutos dos resíduos padronizados em termos do mínimo, mediana e máximo, respectivamente.

8. Por fim, disponha em um gráfico os valores obtidos nos passos 3 e 7 versus os valores esperados da estatística de ordem meio normal calculados pela Equação 4.16.

5 RESULTADOS

Neste Capítulo são apresentados os resultados de três estudos de simulação, além da análise dos dados apresentados no Capítulo 2. O primeiro estudo de simulação foi conduzido para investigar o comportamento do algoritmo NORTA (NOR*mal To Anything*) na simulação de variáveis aleatórias beta correlacionadas (subseção 5.1.1). O segundo visou checar propriedades dos estimadores para os parâmetros de dispersão, no contexto de análise de dados longitudinais (subseção 5.1.2). E o terceiro foi delineado para explorar a flexibilidade dos estimadores para lidar com múltiplas respostas correlacionadas (subseção 5.1.3). Por fim, a subseção 5.2.1 apresenta os resultados da análise dos dados referente ao índice de qualidade da água (IQA), enquanto a subseção 5.2.2 apresenta os resultados correpondentes ao percentual de gordura corporal.

5.1 ESTUDOS DE SIMULAÇÃO

5.1.1 Comportamento do algoritmo NORTA

Fenômenos aleatórios ocorrem em diversas áreas de pesquisa. Esses fenômenos, podem ser representados por uma variável aleatória, sendo descritos por uma distribuição de probabilidades. Para o caso de duas ou mais variáveis aleatórias, pode-se pensar numa distribuição de probabilidades bivariada ou multivariada, além de uma possível correlação existente entre elas. Porém, um dos problemas é especificar a distribuição conjunta do vetor aleatório (DIAS, 2014).

Para simular vetores aleatórios correlacionados vários métodos foram propostos. Li e Hammond (1975) descrevem um método para gerar vetores aleatórios correlacionados. Na sequência, Cario e Nelson (1997) apresentam o algoritmo NORTA (NOR*mal To Anything*), que é um dos métodos mais populares para simulação de variáveis aleatórias correlacionadas não-gaussianas. Aplicações do método podem ser vistas em Chen (2001) e Ghosh e Henderson (2003). Adicionalmente, o algoritmo NORTA foi combinado com o método de redes neurais artificiais para gerar vetores aleatórios correlacionados com distribuições marginais de qualquer tipo, quando se tem problemas em resolver equações não-lineares por métodos convencionais (NIAKI; ABBASI, 2008). Recentemente, Bonat (2017) adaptou o algoritmo NORTA para simular vetores aleatórios correlacionados na análise de dados genéticos.

Fazendo-se uso de estudos de simulação, pode-se investigar o comportamento das variáveis aleatórias, uma vez que se tem o controle sobre o mecanismo gerador dos dados. Além disso, pode-se estudar propriedades dos estimadores, tais como viés,

consistência e taxa de cobertura (ver subseção 5.1.2 e subseção 5.1.3).

Para o caso da distribuição beta bivariada, o principal problema está na identificação dos valores mínimo e máximo que a matriz de correlação pode assumir em função das suas médias marginais. A distribuição beta, na sua forma reparametrizada (FERRARI; CRIBARI-NETO, 2004), é modelada por dois parâmetros: um de média e outro de dispersão denotados, respectivamente, por μ e ϕ. Essa distribuição não é ortogonal, como é o caso da distribuição Gaussiana. Assim, seus parâmetros são definidos por $p = \mu\phi$ e $q = (1 - \mu)\phi$. Para detalhes, ver seção 3.2.

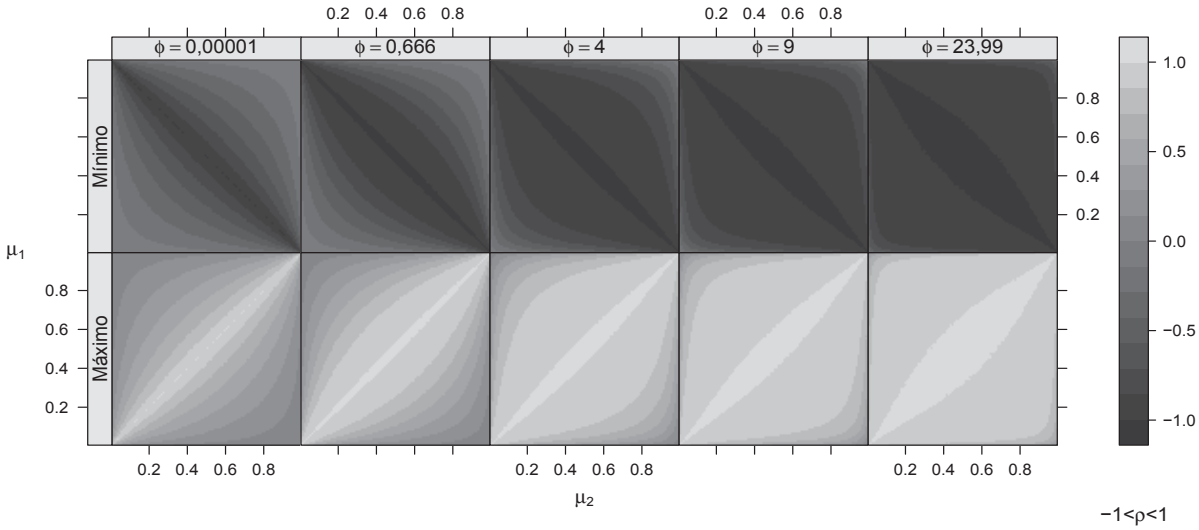
Nesse contexto, o principal objetivo desta seção é estudar o comportamento do algoritmo NORTA para simular variaveis aleatórias beta bivariada. Assim, fez-se uso do *software* estatístico `R` (R DEVELOPMENT CORE TEAM, 2018) e do pacote `NORTARA` (SU, 2014), que fornece toda a implementação computacional do algoritmo NORTA.

O delineamento do estudo de simulação foi feito usando duas variáveis aleatórias beta, denotadas por Y_1 e Y_2, fixando-se diferentes valores para os parâmetros de dispersão $\phi = (0{,}00001;\ 0{,}666;\ 4;\ 9;\ 23{,}99)$. Para cada distribuição marginal foi gerado uma sequência com 100 valores entre $(0,1)$. Assim, construiu-se um grid de valores com 10.000 pontos (100×100) para avaliação da matriz de correlação entre as marginais beta. Como o interesse é saber quais são os valores mínimo e máximo que a matriz de correlação pode assumir entre duas variáveis aleatórias beta, de acordo com suas médias marginais e valores de ϕ, fez-se uso da função `valid_input_cormat()` do pacote `NORTARA`. Esta função retorna os valores mínimo (ρ_L) e máximo (ρ_U) que a matriz de correlação pode assumuir em função das distribuições marginais.

Por exemplo, no caso da distribuição beta, quando $\mu_1 = 0{,}495$ e $\mu_2 = 0{,}851$ com $\phi = 0{,}00001$ (fixo) tem-se $\rho_L = -0{,}421$ e $\rho_U = 0{,}413$, enquanto que para os mesmos valores das médias marginais, porém com $\phi = 9$ (fixo), tem-se $\rho_L = -0{,}956$ e $\rho_U = 0{,}954$. Isso mostra como o parâmetro de dispersão afeta diretamente a média das distribuições marginais e, consequentemente, os limites da matriz de correlação.

Com o objetivo de tornar essa ideia mais geral, construiu-se a Figura 10. A parte superior desta figura mostra os limites mínimos, enquanto a parte inferior da figura apresenta os limites máximos que a correlação entre duas variáveis aleatórias beta assume em função das suas médias marginais para cada valor de ϕ. De acordo com os resultados apresentados na Figura 10, quando se tem valores baixos para ϕ a correlação mínima obtida fica restrita, especialmente quando $\phi = (0{,}00001$ e $0{,}666)$. Na medida em que o valor de ϕ aumenta, correlações mais fortes são observadas nas regiões mais escuras do gráfico. O mesmo é observado para os valores máximos da correlação (Figura 10). Dessa forma, observou-se que altos valores do parâmetro de dispersão combinados com altos/baixos valores das médias marginais produzem correlações mais fortes.

FIGURA 10 – VALORES MÍNIMOS E MÁXIMOS PARA A CORRELAÇÃO ENTRE DUAS VARIÁVEIS ALEATÓRIAS BETA EM FUNÇÃO DAS MÉDIAS MARGINAIS E DIFERENTES VALORES DO PARÂMETRO ϕ



FONTE: O autor (2018).

Os resultados do estudo de simulação mostraram que o espaço paramétrico da correlação ficou reduzido, quando se tem baixos valores para os parâmetros de dispersão associados com altos/baixos valores das médias marginais. Diante dos resultados obtidos, tem-se uma ideia do comportamento do algoritmo NORTA que será usado nos estudos de simulação apresentados na subseção 5.1.2 e subseção 5.1.3 para avaliar o desempenho do método de estimação proposto para os parâmetros do modelo de regressão quase-beta multivariado.

Cabe observar, que a simulação de uma distribuição beta multivariada é uma tarefa desafiadora devido a certas restrições. Primeiro, o suporte das distribuições marginais é o intervalo unitário. Segundo, na distribuição beta existe um relacionamento entre média e variância. Finalmente, dependendo da especificação das distribuições marginais pode ser difícil simular vetores aleatórios dessa distribuição, especialmente quando se tem baixos valores para os parâmetros de dispersão.

5.1.2 Propriedades dos estimadores em estudos longitudinais

O segundo estudo de simulação foi delineado para verificar a robustez dos estimadores propostos na seção 4.2, para lidar com dados limitados em estudos longitudinais. Nesse caso, o principal interesse é avaliar viés e consistência nos parâmetros de dispersão. Para isso, foram criados 36 cenários de simulação levando em conta diferentes estruturas de covariância. Os cenários foram delineados combinando três estruturas de covariância, associadas a três níveis de correlação cada, com quatro valores do

parâmetro de dispersão $\phi = (0{,}666;\ 4;\ 9;\ 23{,}99)$ da densidade beta com μ_{ri} fixa em 0,5. Assim, o preditor linear da estrutura da média foi especificado por $g(\mu_{ri}) = \beta_0$, onde $g(\cdot)$ é a função de ligação *logit* e β_0 é o termo constante (intercepto). Além disso, o preditor linear matricial de cada estrutura de covariância foi especificado pela combinação da matriz identidade com matrizes conhecidas que definem a estrutura de interesse.

Neste estudo de simulação, considerou-se as estruturas de covariância permutável (do inglês *exchangeable*, Exch), médias móveis de ordem 1 (MA1) e uma estrutura baseada em distâncias (Dist^2), que é definida pela matriz de distâncias (Dist) somada a outra matriz de distâncias com seus elementos elevados ao quadrado. Ainda, considerou-se $R = 5$ e $R = 10$, isto é, cinco e dez medidas repetidas no mesmo indivíduo.

Para descrever os componentes do preditor linear matricial para cada caso, suponha, por simplicidade, $R = 3$. Logo, a especificação do preditor linear matricial para cada uma das estruturas acima mencionadas é dada por:

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

no caso permutável, por

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} . & 1 & . \\ 1 & . & 1 \\ . & 1 & . \end{bmatrix},$$

no caso da estrutura MA1, e por

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} . & 1/d_{12} & 1/d_{13} \\ 1/d_{12} & . & 1/d_{23} \\ 1/d_{13} & 1/d_{23} & . \end{bmatrix} + \tau_2 \begin{bmatrix} . & 1/d_{12}^2 & 1/d_{13}^2 \\ 1/d_{12}^2 & . & 1/d_{23}^2 \\ 1/d_{13}^2 & 1/d_{23}^2 & . \end{bmatrix},$$

para a estrutura baseada em distâncias (Dist^2).

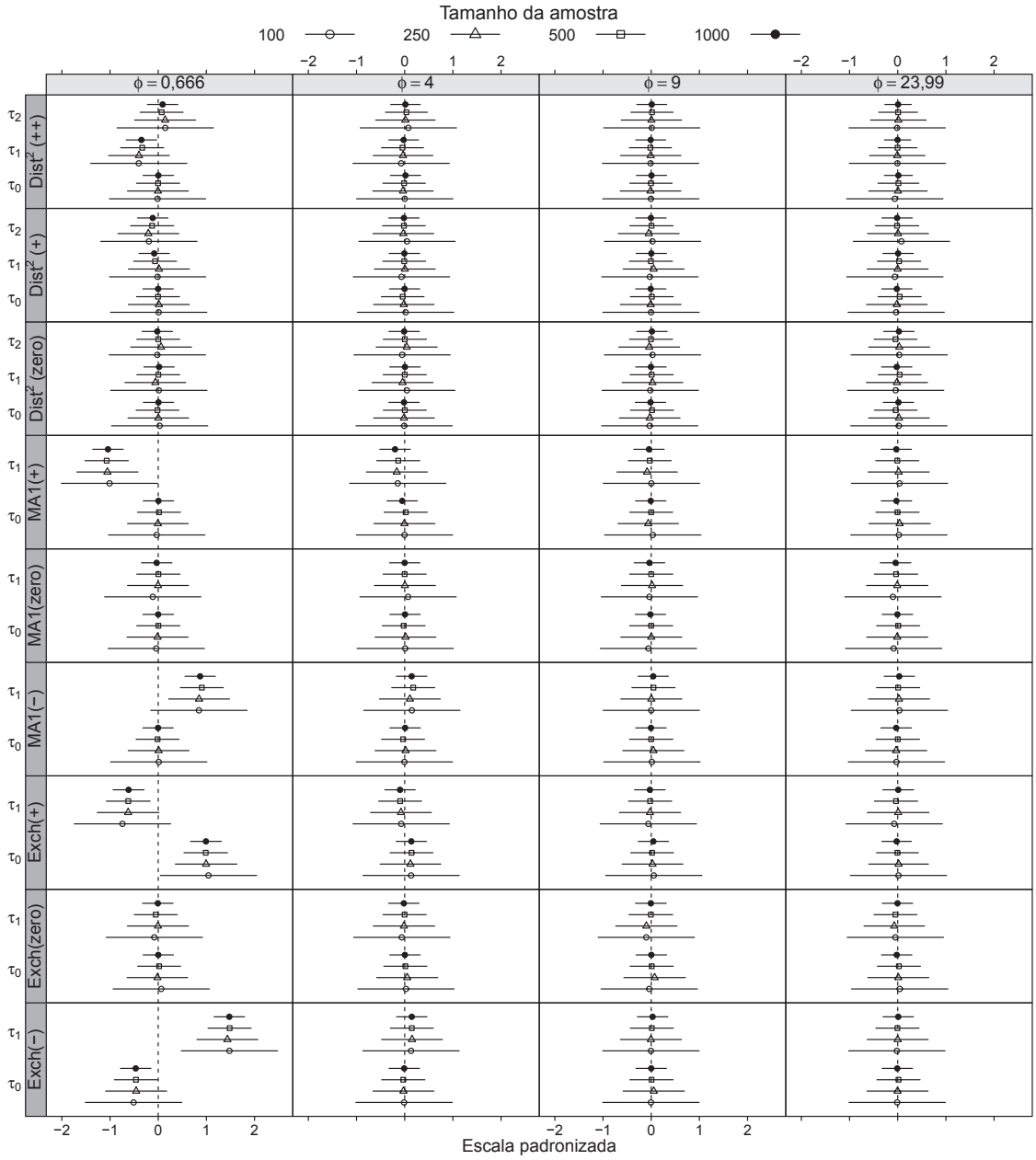
Para as estruturas de covariância Exch e MA1, foram definidas correlações negativa, nula e positiva, enquanto que para a estrutura baseada em distâncias foram consideradas correlações nula, positiva fraca e positiva forte. Portanto, para cada um dos 36 cenários de simulação, foram gerados 300 conjuntos de dados com quatro tamanhos de amostra cada $(100, 250, 500$ e $1000)$. Para simular variáveis aleatórias beta correlacionadas, fez-se uso do algoritmo NORTA (NOR*mal To Anything*), disponível no pacote `NORTARA` (SU, 2014) do *software* estatístico `R` (R DEVELOPMENT CORE TEAM, 2018).

No caso da estrutura de covariância Exch com correlação negativa, definiu-se $\tau_0 = 1{,}2$ e $\tau_1 = -0{,}2$ como os verdadeiros valores dos parâmetros de dispersão, enquanto

que para correlações nula e positiva os parâmetros foram fixados em $(\tau_0 = 1;\ \tau_1 = 0)$ e $(\tau_0 = 0{,}3;\ \tau_1 = 0{,}7)$, respectivamente. Para a estrutura de covariância MA1, com correlações negativa, nula e positiva os parâmetros de dispersão foram fixados em $(\tau_0 = 1;\ \tau_1 = -0{,}5)$, $(\tau_0 = 1;\ \tau_1 = 0)$ e $(\tau_0 = 1;\ \tau_1 = 0{,}5)$, respectivamente. Finalmente, os parâmetros de dispersão $(\tau_0 = 1;\ \tau_1 = 0;\ \tau_2 = 0)$, $(\tau_0 = 1;\ \tau_1 = 0{,}1;\ \tau_2 = 0{,}2)$ e $(\tau_0 = 1;\ \tau_1 = 0{,}25;\ \tau_2 = 0{,}45)$ foram fixados na avaliação da estrutura baseada em distâncias (Dist^2) para os casos com correlações nula, positiva fraca e positiva forte, respectivamente.

A Figura 11 apresenta o viés médio mais ou menos o erro padrão médio para os parâmetros de dispersão sob cada cenário de simulação. Nesta figura, as escalas foram padronizadas para cada parâmetro dividindo-se o viés médio e os limites dos intervalos de confiança pelo erro padrão obtido na amostra de tamanho 100. Além disso, esta figura apresenta os resultados do estudo de simulação para $R = 5$ medidas repetidas. De acordo com os resultados apresentados na Figura 11, os parâmetros τ_1 são viciados sob todos os cenários de simulação nos quais $\phi = 0{,}666$, com exceção dos cenários onde a correlação é nula. Porém, para os outros cenários de simulação os estimadores propostos para os parâmetros de dispersão são não viciados e consistentes, conforme aumenta-se o tamanho da amostra.

FIGURA 11 – VIÉS MÉDIO E INTERVALOS DE CONFIANÇA EM ESCALA PADRONIZADA PARA OS PARÂMETROS DE DISPERSÃO POR TAMANHO DE AMOSTRA E ESTRUTURA DE COVARIÂNCIA COM DIFERENTES NÍVEIS DE CORRELAÇÃO



FONTE: O autor (2018).

No Apêndice B encontra-se uma figura semelhante a Figura 11, que foi construída com $R = 10$ medidas repetidas. Para a construção desta figura algumas configurações foram alteradas no estudo de simulação. Na estrutura de covariância Exch com correlação negativa, fixou-se $\tau_0 = 1,1$ e $\tau_1 = -0,1$ como os verdadeiros valores dos parâmetros de dispersão, e na estrutura Dist^2 com correlação positiva forte, trocou-se apenas um valor, de $\tau_2 = 0,45$ para $\tau_2 = 0,4$. Tais modificações foram necessárias para garantir que a matriz de correlação usada no algoritmo NORTA seja positiva definida.

5.1.3 Propriedades dos estimadores em estudos com múltiplas respostas

Nesta subseção, serão apresentados os resultados do terceiro estudo de simulação, conduzido para investigar as propriedades dos estimadores propostos na seção 4.2 para lidar com dados limitados em estudos com múltiplas respostas correlacionadas. Diferentemente do estudo de simulação apresentado na subseção 5.1.2, este estudo visa verificar viés, consistência e taxa de cobertura nos parâmetros de regressão e dispersão, respectivamente. Para tanto, foram criados 35 cenários de simulação com dados gerados por uma distribuição beta bivariada. Desse modo, os cenários foram delineados pela combinação de cinco valores do parâmetro de dispersão $\phi = (0{,}00001;\ 0{,}666;\ 4;\ 9;\ 23{,}99)$ com sete valores do coeficiente de correlação $\rho = (-0{,}75;\ -0{,}50;\ -0{,}25;\ 0{,}00;\ 0{,}25;\ 0{,}50;\ 0{,}75)$. Tais correlações foram fixadas para supor diferentes graus de depêndencia entre as distribuições marginais beta. Para cada cenário, foram gerados 1000 conjuntos de dados com quatro tamanhos de amostra cada $(100{,}250{,}500$ e $1000)$.

Para simular variáveis aleatórias beta correlacionadas fez-se uso do algoritmo NORTA (CARIO; NELSON, 1997). Assim, o estudo de simulação foi conduzido no *software* estatístico `R` (R DEVELOPMENT CORE TEAM, 2018) com auxílio do pacote `NORTARA` (SU, 2014).

Nas simulações, foram considerados os seguintes preditores lineares:

$$\begin{aligned} g(\mu_{1i}) &= \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i}, \\ g(\mu_{2i}) &= \beta_{02} + \beta_{12}x_{1i} + \beta_{22}x_{2i},\ i = 1,2,\ldots,n, \end{aligned} \tag{5.1}$$

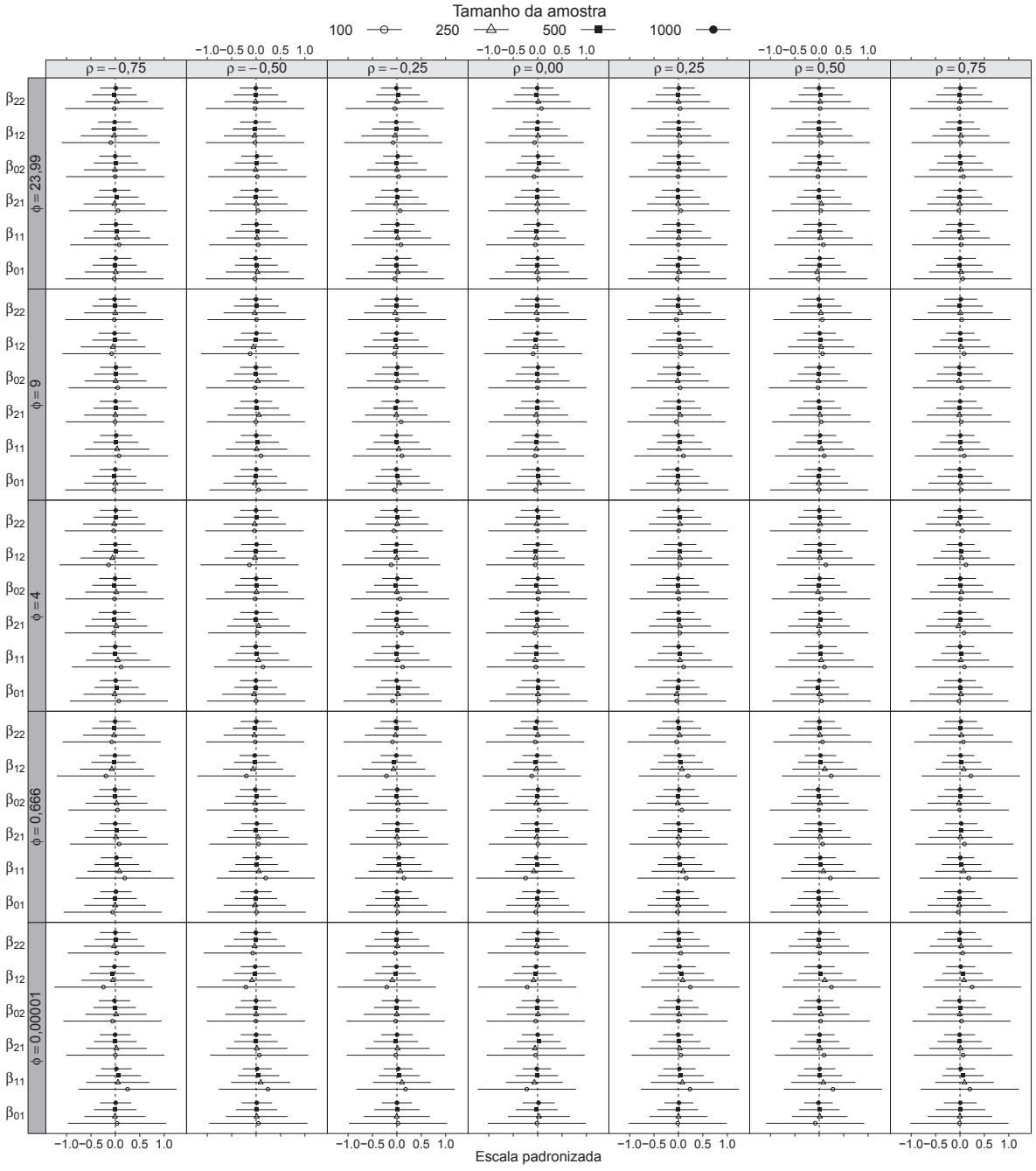
onde $g(\cdot)$ é a função de ligação *logit*. Os preditores lineares definidos na Equação 5.1 foram usados em todos os cenários de simulação e são compostos por uma covariável contínua e outra categórica. A covariável contínua x_1 foi gerada a partir de uma distribuição Gaussiana com média zero e variância fixa em $0{,}3^2$. Já a covariável categórica x_2 foi gerada a partir de uma distribuição Bernoulli com probabilidade fixa em 0,5. Os coeficientes de regressão $(\beta_{01},\ \beta_{11},\ \beta_{21},\ \beta_{02},\ \beta_{12},\ \beta_{22})$ foram fixados de acordo com as correlações assumidas entre as variáveis respostas. Com base no primeiro estudo de simulação (subseção 5.1.1) definiu-se três conjuntos de coeficientes de regressão. Para correlações negativas, os coeficientes foram fixados em $\boldsymbol{\beta} = (0;\ 4;\ 0{,}5;\ 0;\ -4;\ -0{,}5)^\top$, enquanto que para correlações positivas e nula os coeficientes foram fixados em $\boldsymbol{\beta} = (0;\ 4;\ 0{,}5;\ 0;\ 4;\ 0{,}5)^\top$ e $\boldsymbol{\beta} = (0;\ -4;\ -0{,}5;\ 0;\ -4;\ -0{,}5)^\top$, respectivamente. Já os parâmetros $\boldsymbol{\lambda} = (\rho_{12},\ \phi_{11},\ \phi_{12})^\top$ foram fixados de acordo com os cenários de simulação.

A Figura 12 apresenta o viés médio mais ou menos o erro padrão médio para os coeficientes de regressão sob cada cenário, enquanto a Figura 13 apresenta as mesmas informações, porém, para os parâmetros $\boldsymbol{\lambda} = (\rho_{12},\ \phi_{11},\ \phi_{12})^\top$. Apesar destas duas figuras

apresentarem informações semelhantes, elas foram separadas por questões visuais e de espaço, além de permitir que os resultados possam ser melhor interpretados.

É importante observar, que as escalas da Figura 12 e Figura 13 foram padronizadas para cada parâmetro, dividindo-se o viés médio e os limites dos intervalos de confiança pelo erro-padrão obtido na amostra de tamanho 100.

FIGURA 12 – VIÉS MÉDIO E INTERVALOS DE CONFIANÇA EM ESCALA PADRONIZADA PARA OS COEFICIENTES DE REGRESSÃO (β_{01}, β_{11}, β_{21}, β_{02}, β_{12}, β_{22}) POR TAMANHO DE AMOSTRA E CENÁRIO DE SIMULAÇÃO



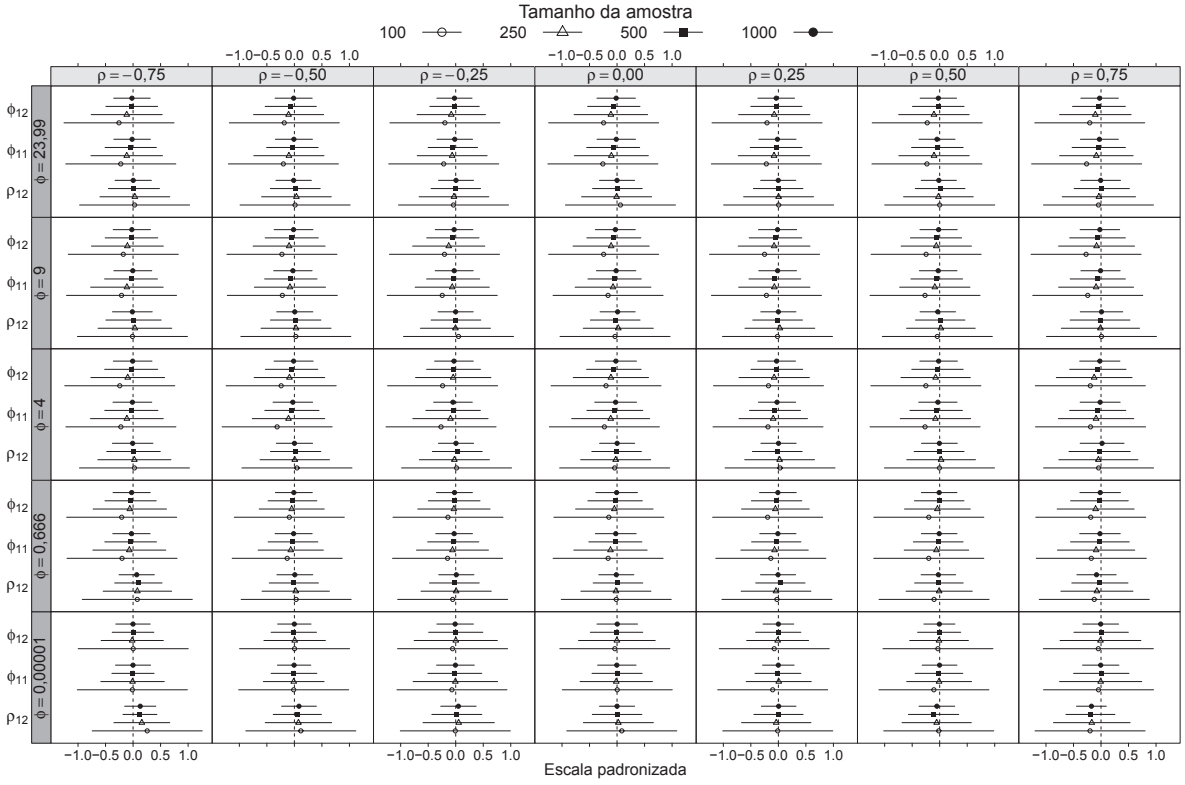
FONTE: O autor (2018).

Os resultados apresentados na Figura 12 mostram que do cenário mais difícil de simular $(\phi = 0{,}00001)$ para o mais fácil $(\phi = 23{,}99)$, independente da correlação que foi fixada, ambos os parâmetros β_{11} e β_{12} são ligeiramente viesados para amostras pequenas. Por outro lado, os parâmetros β_{01}, β_{21}, β_{02} e β_{22} mostraram-se não viciados sob todos os cenários de simulação e tamanhos de amostra.

Portanto, os resultados apresentados na Figura 12 mostram que em todos os cenários de simulação ambos viés e erro-padrão médio tendem a zero à medida que o tamanho da amostra aumenta, mostrando que os estimadores propostos na seção 4.2 para os parâmetros de regressão são consistentes e não-viciados.

Os resultados apresentados na Figura 13 correspondem aos parâmetros de dispersão e correlação, por cenário de simulação e tamanho de amostra.

FIGURA 13 – VIÉS MÉDIO E INTERVALOS DE CONFIANÇA EM ESCALA PADRONIZADA PARA CADA PARÂMETRO $(\rho_{12}, \phi_{11}, \phi_{12})$ POR TAMANHO DE AMOSTRA E CENÁRIO DE SIMULAÇÃO



FONTE: O autor (2018).

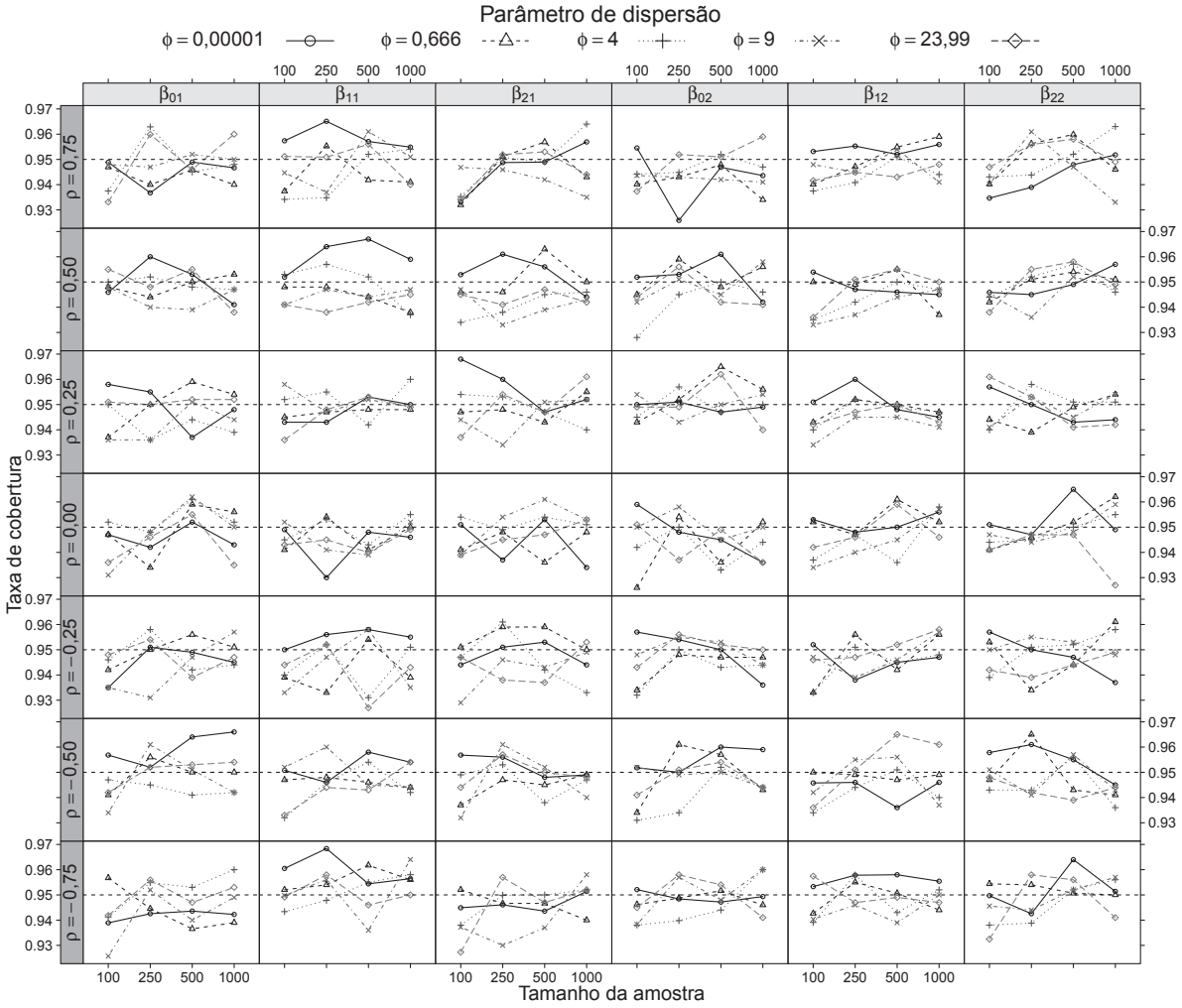
Assim, a Figura 13 mostra que o viés médio e o erro-padrão médio para os parâmetros ϕ_{11} e ϕ_{12} tendem a zero à medida que o tamanho da amostra aumenta. Nos cenários onde $\phi = 0{,}00001$, independente da correlação que foi fixada, os resultados para esses parâmetros foram satisfatórios. No entanto, nos cenários onde $\phi = 0{,}666$ até $\phi = 23{,}99$ os estimadores são viesados para amostras pequenas.

Ainda, de acordo com os resultados apresentados na Figura 13, é possível verificar que o parâmetro de correlação ρ_{12} foi viesado nos cenários onde $\phi = 0{,}00001$, principalmente quando se tem correlações mais extremas $(-0{,}75$ e $0{,}75)$. Porém, nos outros cenários de simulação, ambos viés e erro-padrão médio de ρ_{12} tendem a zero, conforme aumenta-se o tamanho da amostra.

De maneira geral, os estimadores propostos na seção 4.2 para os parâmetros de dispersão e correlação são consistentes e não-viciados para amostras grandes.

A Figura 14 apresenta a taxa de cobertura dos intervalos de confiança para os coeficientes de regressão. Desta figura, observa-se que em todos os cenários de simulação e tamanhos de amostra, as taxas de cobertura dos intervalos de confiança ficaram próximas do nível nominal de 95%.

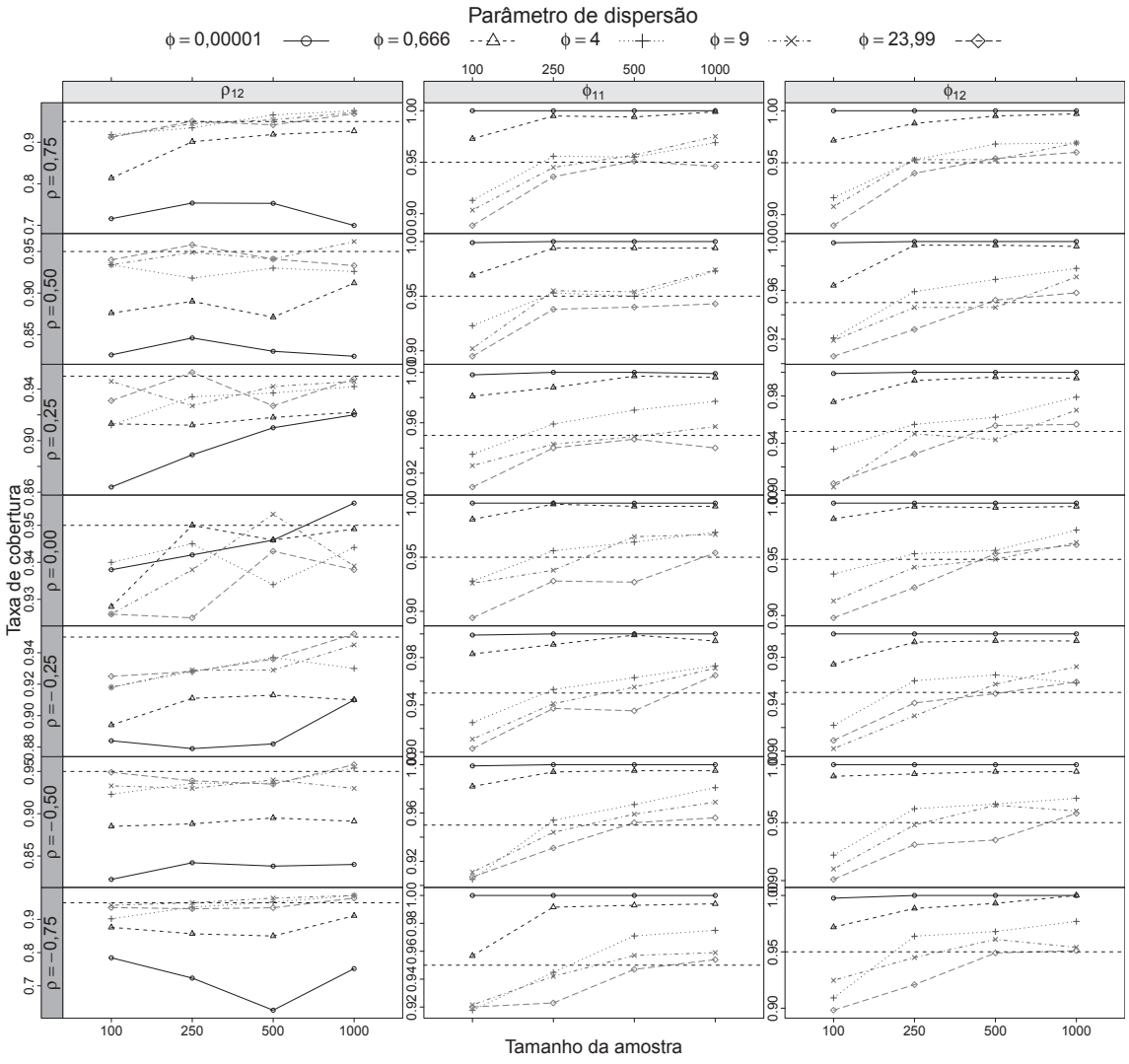
FIGURA 14 – TAXA DE COBERTURA PARA CADA PARÂMETRO $(\beta_{01},\ \beta_{11},\ \beta_{21},\ \beta_{02},\ \beta_{12},\ \beta_{22})$, POR TAMANHO DE AMOSTRA, PARÂMETRO DE DISPERSÃO E COEFICIENTE DE CORRELAÇÃO



FONTE: O autor (2018).

Na Figura 15 encontram-se os resultados da taxa de cobertura para os parâmetros de dispersão (ϕ_{11} e ϕ_{12}) e correlação (ρ_{12}), por tamanho de amostra e cenário de simulação.

FIGURA 15 – TAXA DE COBERTURA PARA CADA PARÂMETRO (ρ_{12}, ϕ_{11}, ϕ_{12}) POR TAMANHO DE AMOSTRA, PARÂMETRO DE DISPERSÃO E COEFICIENTE DE CORRELAÇÃO



FONTE: O autor (2018).

De acordo com os resultados apresentados na Figura 15, a taxa de cobertura para o coeficiente de correlação ficou em torno do nível nominal de 90%, para os cenários com correlações fixas em $-0,25$ e $0,25$, independente do valor fixado para a dispersão. Contudo, para os cenários com correlações mais extremas $(-0,75$ e $0,75)$ o nível nominal dos intervalos de confiança ficaram próximos de 70%, quando $\phi = 0,00001$.

Ainda, conforme mostra a Figura 15, a taxa de cobertura dos parâmetros de dispersão (ϕ_{11} e ϕ_{12}), sob todos os cenários de simulação e tamanhos de amostra, ficaram em torno do nível nominal de 95%.

5.2 RESULTADO DA ANÁLISE DOS DADOS

5.2.1 Análise do índice de qualidade da água

Nesta subseção, são apresentados os principais resultados da análise dos dados da seção 2.1. Os dados se referem ao índice de qualidade da água (IQA), de reservatórios de usinas hidrelétricas operadas pela COPEL no Estado do Paraná. O objetivo da análise é investigar o relacionamento do IQA com a covariável local (montante, reservatório e jusante) controlado pelo efeitos dos trimestres e das usinas. Portanto, o principal desafio deste conjunto de dados está na avaliação de uma variável resposta limitada ao intervalo $(0,1)$, levando em conta características de um estudo longitudinal e de dados agrupados.

Assim, para investigar tais questões, pretende-se modelar a estrutura de covariância $\boldsymbol{\Omega}(\boldsymbol{\tau})$ do modelo proposto no Capítulo 4, com o intuito de investigar possíveis correlações intraunidades amostrais. Em outras palavras, será investigada a presença de correlações entre os locais e os trimestres dentro de cada unidade amostral (usina).

Considere $\mathbf{Y}_i = (Y_{1i}, \ldots, Y_{12i})^\top$ um vetor de variáveis respostas com 12 medidas (3 locais $\times$ 4 trimestres) associado ao índice de qualidade da água da i-ésima usina, para $i = 1, \ldots, 16$. Nessa notação, $r = j \times k$, para $j = 1,2,3$ locais (montante, reservatório e jusante) e $k = 1, \ldots, 4$ trimestres. Denote $\boldsymbol{\mu}_i = (\mu_{1i}, \ldots, \mu_{12i})^\top$ seu respectivo vetor de médias e seja $g(\mu_{jki})$ o preditor linear relacionado ao local j, trimestre k e usina i. Logo, sua representação é dada por:

$$g(\mu_{jki}) = \beta_0 + \beta_{1j}\, \texttt{local}_{ji} + \beta_{2k}\, \texttt{trimestre}_{ki}, \tag{5.2}$$

onde $g(\cdot) : (0,1) \mapsto \mathbb{R}$ é uma função de ligação para dados limitados, β_0 define o termo constante (intercepto) e β_{1j}, para $j = 2$ e 3, avalia mudanças da montante para o reservatório e da montante para a jusante, respectivamente. Já os coeficientes β_{2k}, para $k = 2$, 3 e 4, medem as diferenças do trimestre 1 para os demais trimestres. Note que, as categorias de referência foram definidas para o local montante e trimestre 1.

Nesse contexto, o modelo de regressão quase-beta multivariado fica representado da seguinte forma:

$$\begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{12i} \end{pmatrix} \sim \bullet \left[\begin{pmatrix} \boldsymbol{\mu}_{1i} \\ \vdots \\ \boldsymbol{\mu}_{12i} \end{pmatrix} ; \boldsymbol{\Sigma}_i \right], \; i = 1, \ldots, 16,$$

onde $\boldsymbol{\Sigma}_i = \mathrm{V}(\boldsymbol{\mu}_i)^{\frac{1}{2}} \boldsymbol{\Omega}(\boldsymbol{\tau}) \mathrm{V}(\boldsymbol{\mu}_i)^{\frac{1}{2}}$ é uma matriz 12×12. No contexto de medidas repetidas e análise de dados longitudinais, também existe o interesse em modelar a matriz $\boldsymbol{\Omega}(\boldsymbol{\tau})$, que é a parte da covariância que não depende da estrutura da média. Dessa forma,

foram propostas quatro estruturas de covariância para a matriz $\boldsymbol{\Omega}(\boldsymbol{\tau})$. A primeira estrutura supõem independência entre as observações, sendo composta por uma matriz identidade. É importante destacar, que as três estruturas apresentadas a seguir são compostas por uma matriz identidade além de outras matrizes que definem a estrutura de interesse. A segunda estrutura é conhecida por permutável (do inglês *exchangeable*), sendo definida por uma matriz constituída de 1's, conforme Equação 5.3.

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{5.3}$$

A matriz não estruturada, usada para avaliar o efeito dos locais (dados agrupados) é dada por

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} . & 1 & . \\ 1 & . & . \\ . & . & . \end{bmatrix} + \tau_2 \begin{bmatrix} . & . & 1 \\ . & . & . \\ 1 & . & . \end{bmatrix} + \tau_3 \begin{bmatrix} . & . & . \\ . & . & 1 \\ . & 1 & . \end{bmatrix}, \tag{5.4}$$

onde τ_1 mede a covariância entre a montante e o reservatório e, τ_2 e τ_3 avaliam a covariância entre a montante e a jusante e, entre o reservatório e a jusante, respectivamente. Logo, a terceira estrutura, chamada por não estruturada 1, é composta pela combinação da estrutura 2 com a especificação apresentada na Equação 5.4.

Para avaliação do efeito dos trimestres (dados longitudinais), a matriz não estruturada tem a seguinte representação

$$\begin{aligned} \boldsymbol{\Omega}(\boldsymbol{\tau}) \;=\; & \tau_0 \begin{bmatrix} 1 & . & . & . \\ . & 1 & . & . \\ . & . & 1 & . \\ . & . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} . & 1 & . & . \\ 1 & . & . & . \\ . & . & . & . \\ . & . & . & . \end{bmatrix} + \tau_2 \begin{bmatrix} . & . & 1 & . \\ . & . & . & . \\ 1 & . & . & . \\ . & . & . & . \end{bmatrix} + \tau_3 \begin{bmatrix} . & . & . & 1 \\ . & . & . & . \\ . & . & . & . \\ 1 & . & . & . \end{bmatrix} \\ & + \tau_4 \begin{bmatrix} . & . & . & . \\ . & . & 1 & . \\ . & 1 & . & . \\ . & . & . & . \end{bmatrix} + \tau_5 \begin{bmatrix} . & . & . & . \\ . & . & . & 1 \\ . & . & . & . \\ . & 1 & . & . \end{bmatrix} + \tau_6 \begin{bmatrix} . & . & . & . \\ . & . & . & . \\ . & . & . & 1 \\ . & . & 1 & . \end{bmatrix}. \end{aligned} \tag{5.5}$$

Note que, as covariâncias entre os trimestres são avaliadas pelos coeficientes τ_1 à τ_6. Por exemplo, a covariância entre os trimestres 1 e 2 é avaliada pelo coeficiente τ_1. Já a covariância entre os trimestres 1 e 3 e, entre os trimestres 1 e 4 é avaliada pelos coeficientes τ_2 e τ_3, respectivamente. Para as demais avaliações, a interpretação é feita de forma semelhante. Portanto, a quarta estrutura (não estruturada 2), é definida pela combinação da estrutura 3 com a especificação mostrada na Equação 5.5.

Na sequência, ajustou-se o modelo de regressão quase-beta multivariado aos dados do IQA, considerando as quatro estruturas acima mencionadas além de especificar a função de ligação *logit* para o preditor linear (Equação 5.2).

A Tabela 2 apresenta o valor maximizado do logaritmo da função de pseudo verossimilhança (plogLik), graus de liberdade (df) e os valores dos pseudo critérios de informação de *Akaike* (pAIC) e Bayesiano (pBIC) para o modelo proposto, ajustado sob diferentes estruturas de covariância. A principal diferença entre as quatro estruturas de covariância está na quantidade de parâmetros que são estimados, além da forma em que os dados longitudinais e agrupados são considerados.

TABELA 2 – VALOR MAXIMIZADO DO LOGARITMO DA FUNÇÃO DE PSEUDO VEROSSIMILHANÇA (plogLik), GRAUS DE LIBERDADE (df) E PSEUDO CRITÉRIOS DE INFORMAÇÃO DE AKAIKE (pAIC) E BAYESIANO (pBIC) PARA DIFERENTES ESTRUTURAS DE COVARIÂNCIA

Estrutura	plogLik	df	pAIC	pBIC
Independente	212,06	7	−410,12	−387,39
Permutável	217,06	8	−418,12	−392,14
Não estruturada 1	222,03	11	−422,06	−386,34
Não estruturada 2	234,36	17	−434,72	−379,52

FONTE: O autor (2018).

Os resultados apresentados na Tabela 2 mostram que tanto o valor da função de pseudo verossimilhança ($\texttt{plogLik} = 234,36$) quanto o valor do pseudo critério de informação de *Akaike* ($\texttt{pAIC} = -434,72$) indicam a quarta estrutura (não estruturada 2) como aquela que apresenta o melhor ajuste aos dados. Dessa forma, selecionou-se esta estrutura de covariância para análise dos dados, uma vez que ela é a mais completa. No Apêndice C, encontram-se detalhes sobre a especificação do preditor linear matricial referente a estrutura de covariância selecionada.

A Tabela 3 apresenta os resultados da estatística de Wald, graus de liberdade (df) e p-valores para cada covariável que compõem o preditor linear (Equação 5.2).

TABELA 3 – ESTATÍSTICA DE WALD (W_s), GRAUS DE LIBERDADE (df) E P-VALORES ASSOCIADOS AO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO IQA

Efeitos	W_s	df	p-valor
Local	8,56	2	0,01
Trimestre	8,34	3	0,04

FONTE: O autor (2018).

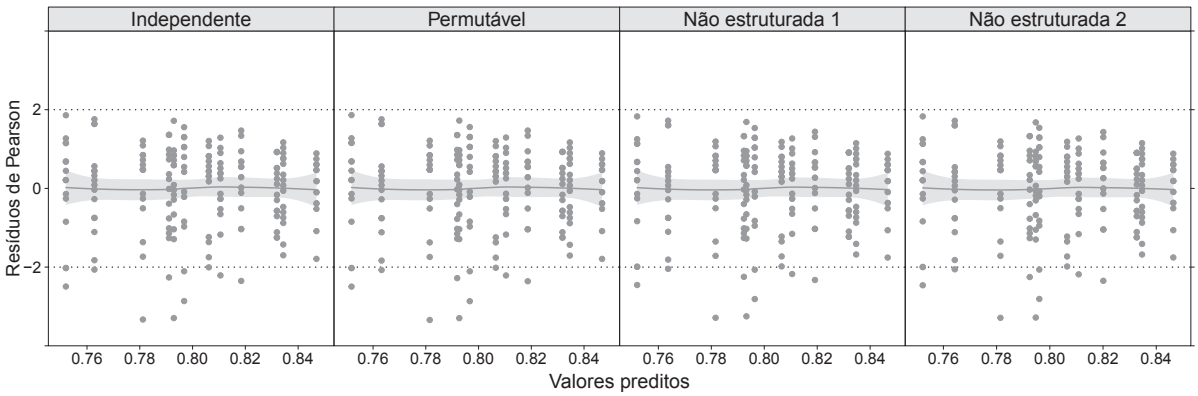
Os resultados na Tabela 3 mostram que ambas as covariáveis são significativamente diferentes de zero e, portanto, são relevantes na análise dos dados.

Após selecionar a estrutura de covariância e analisar o efeito das covariáveis, serão apresentadas a seguir a análise de resíduos e de diagnóstico, afim de avaliar a

qualidade produzida pelo modelo, além de investigar possíveis pontos influentes e *outliers*.

A Figura 16 apresenta os resíduos de Pearson versus os valores preditos pelo modelo de regressão quase-beta multivariado, ajustado por diferentes estruturas de covariância. Além dos resíduos de Pearson, esta figura também mostra curvas de suavização com bandas de confiança estimadas pelo método *loess* (CLEVELAND, 1979).

FIGURA 16 – RESÍDUOS DE PEARSON ASSOCIADOS AO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO ÍNDICE DE QUALIDADE DA ÁGUA (IQA)

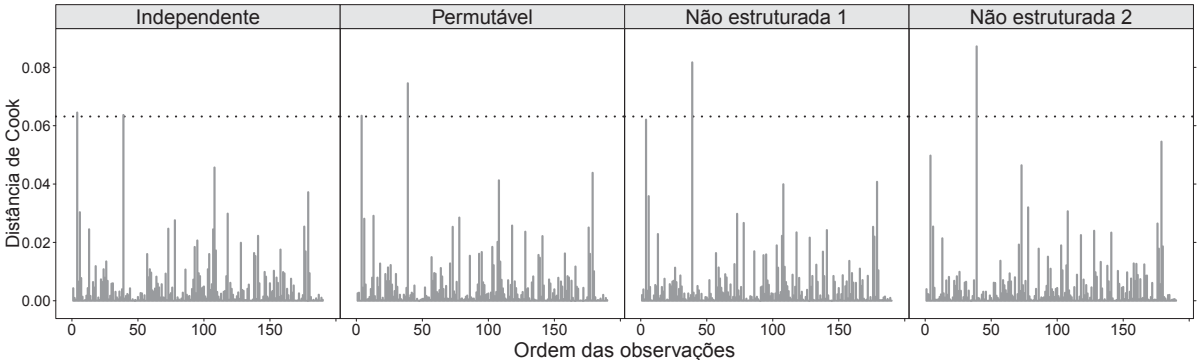


FONTE: O autor (2018).

Os resultados mostrados na Figura 16 apontam que, independente da estrutura de covariância considerada, o modelo proposto apresentou um ajuste satisfatório aos dados do IQA, uma vez que os resíduos variam entre -2 e 2. Apesar de alguns pontos ficarem abaixo do limite inferior, o ajuste por ambos os modelos parecem adequados.

Na sequência, são apresentados os gráficos da distância de Cook versus a ordem das observações (Figura 17) para o modelo proposto ajustado sob cada estrutura de covariância. Para avaliação da distância de Cook, define-se $2p/n$ como ponto de corte, onde p é a quantidade de coeficientes de regressão estimados pelo modelo e n é o tamanho da amostra. Assim, valores acima de $0,063$ são considerados pontos influentes.

FIGURA 17 – DISTÂNCIA DE COOK PARA O MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO ÍNDICE DE QUALIDADE DA ÁGUA (IQA)

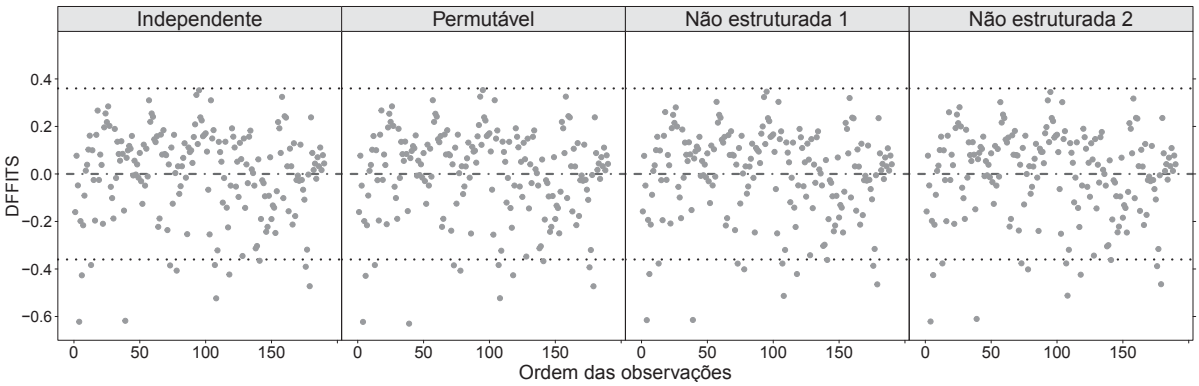


FONTE: O autor (2018).

De acordo com os resultados apresentados na Figura 17, as observações 4 e 39 foram indicadas como pontos influentes no ajuste do modelo sob as estruturas independente e permutável, enquanto que as demais estruturas indicaram apenas a observação 39 como ponto influente.

A Figura 18 apresenta a medida DFFITS para os modelos de regressão ajustados aos dados do IQA. Em geral, essa medida é usada para avaliar a influência da exclusão da i-ésima observação no seu valor estimado pelo modelo. Para a medida DFFITS, o ponto de corte é definido por $2\sqrt{\frac{p}{n}}$, onde p e n se referem a quantidade de coeficientes de regressão e ao tamanho da amostra, respectivamente. Nesse caso, valores fora do intervalo $\pm 0,355$ devem ser investigados. Portanto, com base na medida DFFITS, não há evidências de que o ajuste produzido pelos modelos estejam inadequados.

FIGURA 18 – DFFITS PARA O MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO ÍNDICE DE QUALIDADE DA ÁGUA (IQA)



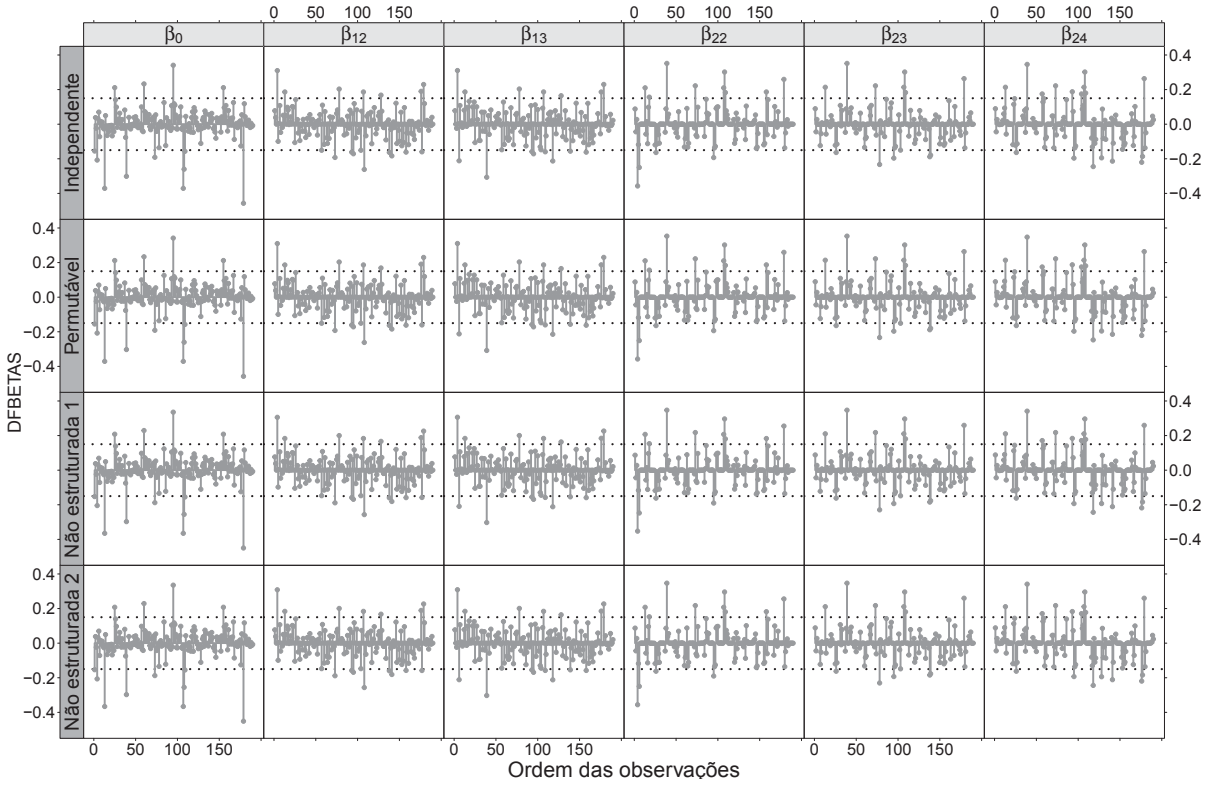
FONTE: O autor (2018).

Os resultados associados a medida DFBETA são mostrados na Figura 19. Tal

medida é comumente usada para avaliar a influência da i-ésima observação sobre cada coeficiente de regressão. Seu ponto de corte é dado por $2/\sqrt{n}$, onde n é o tamanho da amostra. Em particular, para os dados do IQA, valores fora do intervalo $\pm 0,145$ são considerados pontos influentes sobre os parâmetros de regressão.

Conforme os resultados apresentados na Figura 19, existem alguns pontos influentes no ajuste dos modelos de regressão. Além disso, observa-se que esses pontos são os mesmos no ajuste de cada um dos quatro modelos.

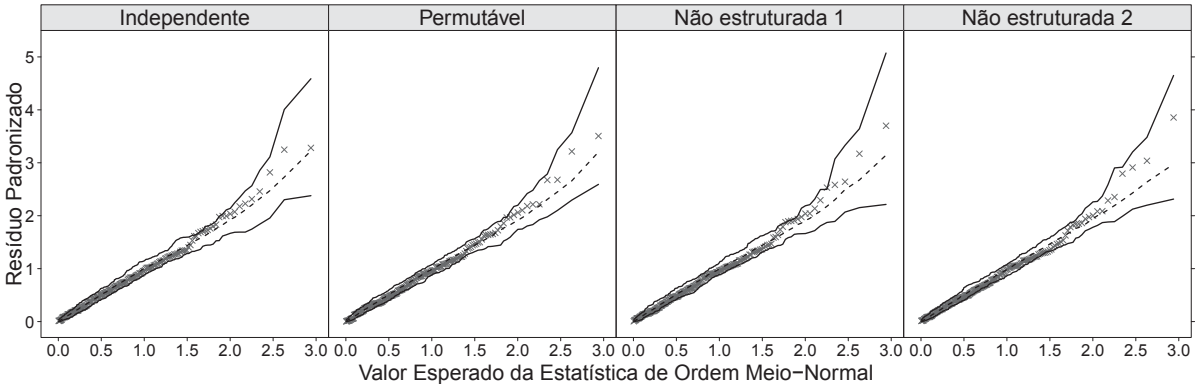
FIGURA 19 – DFBETAS PARA O MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO ÍNDICE DE QUALIDADE DA ÁGUA (IQA)



FONTE: O autor (2018).

A Figura 20 apresenta o gráfico de probabilidade meio-normal com envelope simulado para o modelo proposto ajustado aos dados do IQA, sob cada uma das estruturas de covariância. Essa técnica de diagnóstico é comumente usada para avaliar a qualidade produzida pelo modelo, além de auxiliar na detecção de *outliers*.

FIGURA 20 – GRÁFICO DE PROBABILIDADE MEIO-NORMAL COM ENVELOPE SIMULADO AJUSTADO AOS DADOS DO ÍNDICE DE QUALIDADE DA ÁGUA (IQA) PELO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO USANDO DIFERENTES ESTRUTURAS DE COVARIÂNCIA



FONTE: O autor (2018).

Os resultados apresentados na Figura 20 indicam que poucos pontos ficaram fora das bandas de confiança no ajuste produzido pelos quatro modelos. Tais resultados indicam que o ajuste do modelo de regressão quase-beta multivariado foi adequado sob diferentes estruturas de covariância. Cabe lembrar, que neste procedimento, assumiu-se distribuição de probabilidade beta para as variáveis respostas simuladas. Assim, mostrou-se que o modelo de regressão proposto nesta dissertação fornece uma aproximação adequada para conjuntos de dados gerados por uma distribuição marginal beta.

Em geral, tanto os resíduos de Pearson como as demais medidas (distância de Cook, DFFITS, DFBETAS e gráfico de probabilidade meio-normal) indicam que o modelo de regressão quase-beta multivariado ajustou-se adequadamente aos dados do IQA.

Após a análise de resíduos e de diagnóstico, serão apresentadas as principais interpretações dos parâmetros estimados pelo modelo. Assim, a Tabela 4 mostra as estimativas dos parâmetros de regressão, erros-padrão, razão de chances, p-valores além de outras informações associadas ao modelo de regressão quase-beta multivariado. É importante destacar, que a razão de chances e os intervalos de confiança apresentados na Tabela 4 foram calculados de forma usual, ou seja, de maneira semelhante ao modelo de regressão beta (seção 3.4).

TABELA 4 – ESTIMATIVAS DOS PARÂMETROS DE REGRESSÃO (Est.), ERROS-PADRÃO (EP), RAZÃO DE CHANCES (RC) E INTERVALOS (IC) COM 95% DE CONFIANÇA, Z-VALOR E *P*-VALORES ASSOCIADOS AO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO IQA

Efeito	Est.	EP	Z-valor	RC (IC 95%)	p-valor
β_0: `Intercepto`	1,111	0,103	10,745	$---------$	$< 0,001$
β_{12}: `Reservatório`	0,251	0,098	2,552	1,286 (1,060−1,559)	0,011
β_{13}: `Jusante`	0,163	0,102	1,604	1,178 (0,964−1,438)	0,109
β_{22}: `Trimestre 2`	0,243	0,131	1,848	1,275 (0,985−1,650)	0,065
β_{23}: `Trimestre 3`	0,345	0,132	2,617	1,412 (1,090−1,828)	0,009
β_{24}: `Trimestre 4`	0,065	0,106	0,618	1,068 (0,867−1,315)	0,537

FONTE: O autor (2018).

De acordo com os resultados apresentados na Tabela 4, a razão de chances do IQA no reservatório foi 1,286 vezes a da montante. Durante o trimestre 3, a razão de chances do IQA foi estimada em 1,412 vezes a do trimestre 1. Por outro lado, os valores do IQA para os trimestres 2 e 4 foram semelhantes ao do trimestre 1, uma vez que os parâmetros associados a essas avaliações não apresentaram relevância (p-valor $> 0,05$). Além disso, a diferença do IQA entre a jusante e a montante também não foi importante na análise dos dados (p-valor $= 0,109$). Portanto, os resultados estimados pelo modelo de regressão quase-beta multivariado, são concordantes com a análise descritiva e exploratória apresentada na seção 2.1.

A Tabela 5 apresenta as estimativas dos parâmetros de dispersão e seus respectivos erros-padrão, estatística Z e p-valores. As estruturas associadas aos locais e aos trimestres são modeladas pelos parâmetros (τ_2, τ_3, τ_4) e $(\tau_5, \tau_6, \tau_7, \tau_8, \tau_9, \tau_{10})$, respectivamente. Para a estrutura de dados agrupados, nenhum dos parâmetros apresentou significância ao nível de 5%. Com relação a estrutura longitudinal, apenas os parâmetros τ_5, τ_6 e τ_9 foram significativamente diferentes de zero. Apesar de alguns parâmetros de dispersão não apresentarem relevância, a interpretação relacionada a eles será mostrada a seguir como forma ilustrativa. Dessa forma, o principal interesse nos parâmetros de dispersão é avaliar correlações intraunidades amostrais.

TABELA 5 – ESTIMATIVAS DOS PARÂMETROS DE DISPERSÃO (Est.), ERROS-PADRÃO (EP), Z-VALOR E *P*-VALORES ASSOCIADOS AO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO IQA

Parâmetro	Est.	EP	Z-valor	p-valor
τ_0	0,019	0,003	5,460	$< 0,001$
τ_1	0,020	0,005	3,913	$< 0,001$
τ_2	$-0,007$	0,004	$-1,662$	0,097
τ_3	$-0,008$	0,004	$-1,832$	0,067
τ_4	0,002	0,001	1,697	0,089
τ_5	$-0,015$	0,007	$-2,191$	0,028
τ_6	$-0,015$	0,006	$-2,235$	0,025
τ_7	$-0,008$	0,005	$-1,648$	0,099
τ_8	$-0,011$	0,005	$-1,821$	0,069
τ_9	$-0,016$	0,007	$-2,286$	0,022
τ_{10}	$-0,008$	0,005	$-1,616$	0,106

FONTE: O autor (2018).

Por exemplo, a correlação entre a montante e o reservatório foi estimada em $\hat{\rho}_{12} = 0,3217$, sendo o cálculo obtido por $(\hat{\tau}_1 + \hat{\tau}_2)/(\hat{\tau}_0 + \hat{\tau}_1)$. Assim, para tornar essa ideia mais geral, construiu-se a matriz de correlação para o efeito dos locais, dada por

$$\hat{\boldsymbol{\Omega}}(\boldsymbol{\tau})_{\text{Local}} = \begin{bmatrix} 1 & & \\ 0,3217(0,1025) & 1 & \\ 0,2909(0,1078) & 0,5607(0,0284) & 1 \end{bmatrix}, \tag{5.6}$$

onde os números entre parênteses denotam o erro padrão calculado pelo método delta. Desta matriz, observa-se que a correlação entre o reservatório e a jusante foi estimada em $\hat{\rho}_{23} = 0,5607$. A menor correlação foi entre a montante e a jusante $(\hat{\rho}_{13} = 0,2909)$. Tal resultado é esperado, uma vez que a água do rio passa primeiro pela montante, entra no reservatório saindo no sentido da jusante. Em outras palavras, a distância entre os locais induz correlações mais fortes ou mais fracas.

Na sequência, encontra-se a matriz de correlação para o efeito dos trimestres:

$$\hat{\boldsymbol{\Omega}}(\boldsymbol{\tau})_{\text{Trimestre}} = \begin{bmatrix} 1 & & & \\ 0,1146(0,1604) & 1 & & \\ 0,1281(0,1542) & 0,2367(0,1395) & 1 & \\ 0,2857(0,1278) & 0,1025(0,1591) & 0,2919(0,1264) & 1 \end{bmatrix}. \tag{5.7}$$

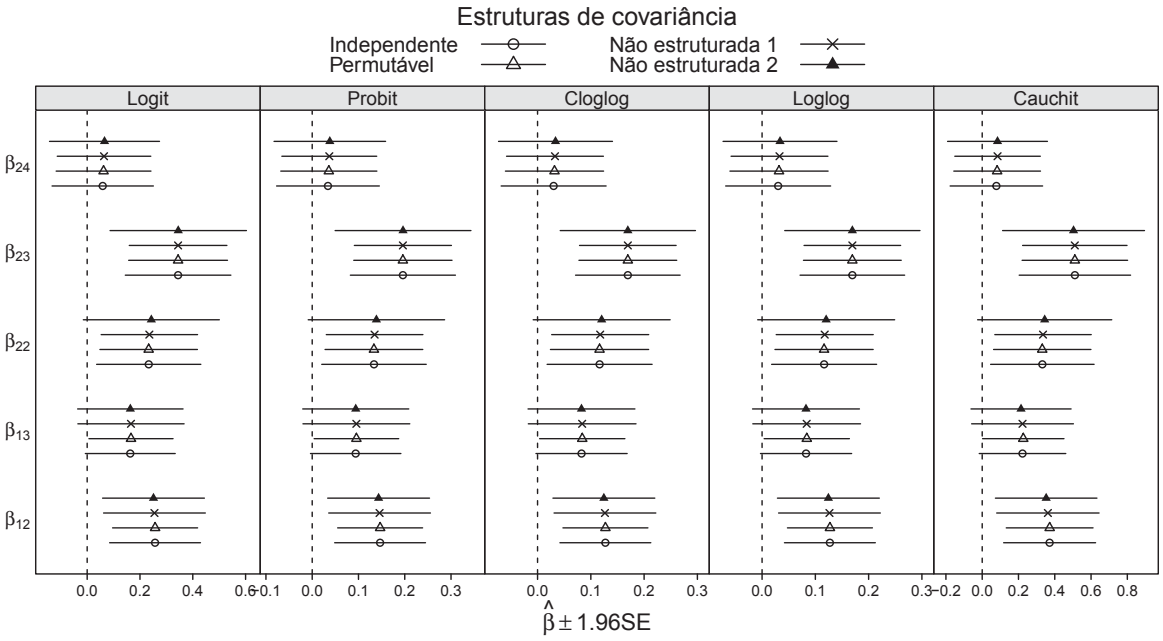
Assim, por exemplo, a correlação entre os trimestres 1 e 2 foi estimada em $\hat{\rho}_{12} = 0,1146$, sendo esta obtida por $(\hat{\tau}_1 + \hat{\tau}_5)/(\hat{\tau}_0 + \hat{\tau}_1)$. Note que, os resultados apresentados na matriz acima mostram correlações fracas entre os trimestres. A correlação entre os

trimestres 1 e 4 foi estimada em $\hat{\rho}_{14} = 0,2857$. Esse resultado é esperado, uma vez os trimestres são cíclicos e, portanto, os trimestres 1 e 4 são próximos.

Na Figura 21 encontram-se os resultados do modelo de regressão quase-beta multivariado ajustado sob diferentes estruturas de covariância e funções de ligação. Os resultados são comparados por meio dos coeficientes de regressão e seus respectivos intervalos com 95% de confiança. O coeficiente β_0 foi omitido para evitar problemas com a escala do gráfico. Logo, a Figura 21 mostra que os resultados obtidos pelas quatro estruturas de covariância são concordantes para todas as funções de ligação. Os coeficientes β_{13} e β_{22} foram os que não apresentaram significância quando avaliados pela estrurura de covariância não estruturada 2, a escolhida para análise dos dados. Além do coeficiente β_{22} não ser significativo, o comprimento do seu intervalo de confiança foi maior na estrutura de covariância usada para análise dos dados, levando em conta todas as funções de ligação avaliadas. Tais resultados, mostram a importância de avaliar diferentes estruturas de covariância, uma vez que as interpretações dos parâmetros variam entre as estruturas.

É importante destacar que, os coeficientes β_{13} e β_{22} medem o efeito da jusante e do trimestre 2, respectivamente. Assim, quando comparam-se as categorias de referência (montante e trimestre 1) com esses coeficientes, seus efeitos são nulos, isto é, o efeito entre os trimestres 1 e 2 são semelhantes, assim como o efeito entre montante e jusante.

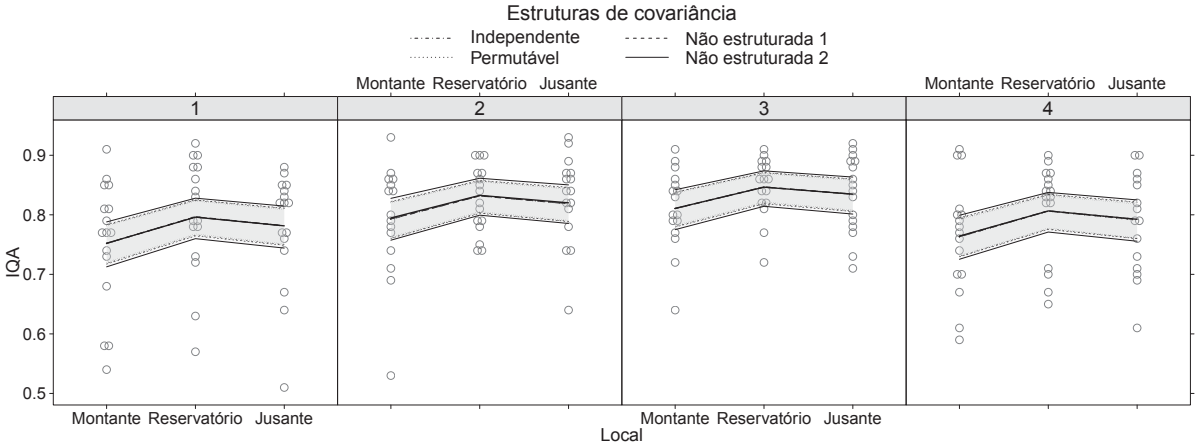
FIGURA 21 – ESTIMATIVAS DOS PARÂMETROS E INTERVALOS COM 95% DE CONFIANÇA PARA OS DADOS DO IQA AJUSTADO PELO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO USANDO DIFERENTES FUNÇÕES DE LIGAÇÃO E ESTRUTURAS DE COVARIÂNCIA



FONTE: O autor (2018).

A Figura 22 apresenta curvas de predição com bandas de confiança (95%) para a média do IQA obtida pelo modelo de regressão quase-beta multivariado, ajustado sob diferentes estruturas de covariâncias.

FIGURA 22 – CURVAS DE PREDIÇÃO COM BANDAS DE CONFIANÇA (95%) PARA A MÉDIA DO IQA POR LOCAL E TRIMESTRE AJUSTADO PELO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO USANDO DIFERENTES ESTRUTURAS DE COVARIÂNCIA

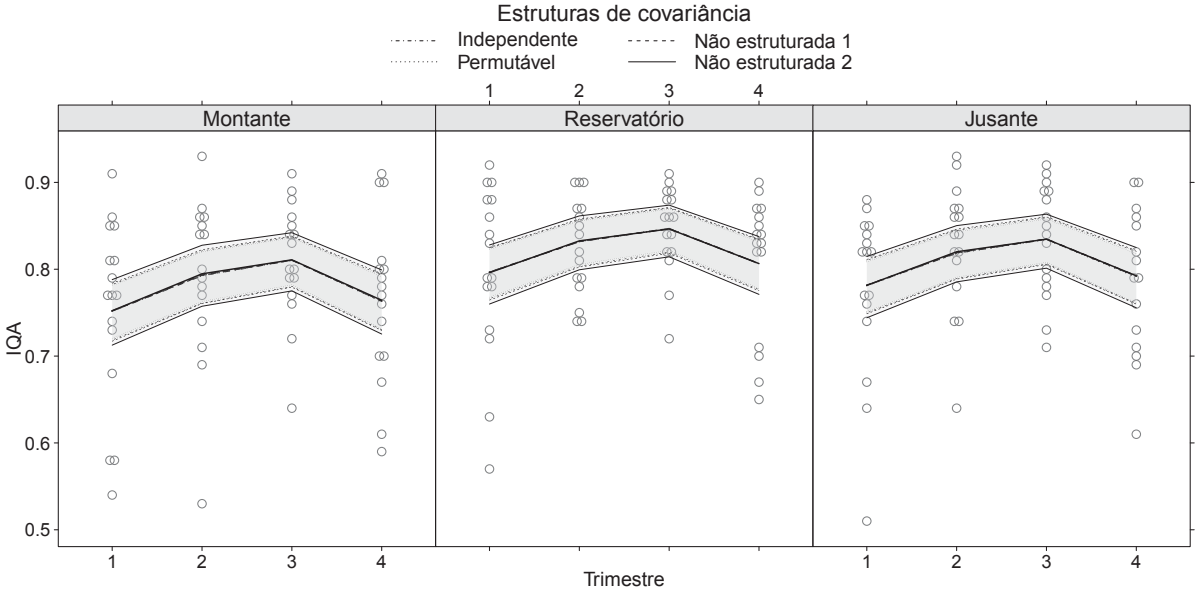


FONTE: O autor (2018).

Os resultados apresentados na Figura 22 mostram pequenas diferenças entre os ajustes, tanto nas estimativas pontuais quanto nos intervalos de confiança. Além disso, os resultados apresentados nesta figura confirmam as hipóteses levantadas na Figura 2 (seção 2.1) e indicam que o IQA foi maior para os dados coletados no terceiro trimestre, bem como no reservatório.

A Figura 23 apresenta os mesmos resultados da figura anterior, porém, de forma alternativa apenas para ilustrar o ajuste do modelo proposto.

FIGURA 23 – CURVAS DE PREDIÇÃO COM BANDAS DE CONFIANÇA (95%) PARA A MÉDIA DO IQA POR TRIMESTRE E LOCAL AJUSTADO PELO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO USANDO DIFERENTES ESTRUTURAS DE COVARIÂNCIA



FONTE: O autor (2018).

Portanto, esta subseção apresentou uma análise de dados com características de um estudo longitudinal e de dados agrupados, mostrando a flexibilidade do modelo proposto nesta dissertação para lidar com dados limitados em estudos com tais características. Na próxima subseção será discutida uma análise diferente desta, onde o conjunto de dados apresenta múltiplas variáveis respostas correlacionadas.

5.2.2 Análise do percentual de gordura corporal

Esta subseção mostra os principais resultados da análise dos dados apresentados na seção 2.2, que correspondem ao percentual de gordura corporal de indivíduos avaliados no HC-UFPR. Tais resultados foram obtidos pelo modelo de regressão quase-beta multivariado proposto no Capítulo 4. Dessa forma, são apresentadas as estimativas dos parâmetros, análise de resíduos e de diagnóstico (distância de Cook, DFBETAS, DFFITS e o gráfico de probabilidade meio-normal com envelope simulado), matriz de correlação entre as respostas além de curvas de predição com bandas de confiança para a média das variáveis respostas.

Como descrito na seção 2.2, este conjunto de dados apresenta cinco variáveis respostas, que se referem aos percentuais de gordura na região dos braços, pernas, tronco, androide e ginecoide (Figura 3). Além disso, o conjunto de dados conta com a presença de quatro covariáveis: sexo, idade, IMC e IPAQ. Assim, o objetivo da análise

dos dados é investigar o relacionamento das cinco variáveis respostas com as covariáveis supracitadas. O modelo descrito no Capítulo 4 deve ser capaz de analisar conjuntamente as cinco variáveis respostas e ainda, estimar a matriz de correlação entre as variáveis respostas levando em conta o efeito das covariáveis.

De acordo com os resultados mostrados na seção 2.2, existe correlação positiva entre as variáveis respostas e há um indicativo de que elas estão relacionadas com as covariáveis do estudo (ver Figura 4 e Figura 5). Com o ajuste do modelo de regressão quase-beta multivariado, espera-se que esses resultados se confirmem, além de esperar que mais informações possam ser obtidas na análise dos dados.

Considere $\mathbf{Y}_i = (Y_{1i}, \ldots, Y_{5i})^\top$ um vetor de variáveis respostas associado, respectivamente, aos percentuais de gordura nos braços, pernas, tronco, androide e ginecoide do indivíduo i, para $i = 1, \ldots, 298$. Seja $\boldsymbol{\mu}_i = (\mu_{1i}, \ldots, \mu_{5i})^\top$ seu respectivo vetor de médias e denote $g_r(\mu_{ri})$ o preditor linear relacionado a r-ésima resposta, para $r = 1, \ldots, 5$, dado por:

$$g_r(\mu_{ri}) = \beta_{0r} + \beta_{1r}\ \texttt{idade}_i + \beta_{2r}\ \texttt{IMC}_i + \beta_{3r}\ \texttt{sexo-M}_i + \beta_{4r}\ \texttt{IPAQ-IA}_i + \beta_{5r}\ \texttt{IPAQ-A}_i, \tag{5.8}$$

onde $g_r(\cdot) : (0,1) \mapsto \mathbb{R}$ é a função de ligação *logit*.

O preditor linear definido na Equação 5.8 será usado em toda análise dos dados, o qual é composto por duas covariáveis contínuas e duas categóricas. As covariáveis contínuas se referem a idade (anos) e IMC (Kg/m^2) dos indivíduos, enquanto as covariáveis categóricas correspondem ao sexo (F-feminino ou M-masculino) e IPAQ (S-sedentário, IA-insuficientemente ativo ou A-ativo). Para as covariáveis categóricas, definiu-se sexo feminino e IPAQ sedentário como as categorias de referência.

Logo, para este particular conjunto de dados, o modelo de regressão quase-beta multivariado fica expresso por:

$$\begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{5i} \end{pmatrix} \sim \bullet \left[\begin{pmatrix} \mu_{1i} \\ \vdots \\ \mu_{5i} \end{pmatrix}; \boldsymbol{\Sigma}_i \right], \ i = 1, \ldots, 298,$$

onde $\boldsymbol{\Sigma}_i = \mathrm{V}(\boldsymbol{\mu}_i)^{\frac{1}{2}} \boldsymbol{\Omega}(\boldsymbol{\lambda}) \mathrm{V}(\boldsymbol{\mu}_i)^{\frac{1}{2}}$ é uma matriz 5×5 com a seguinte forma:

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sqrt{\mu_{1i}(1-\mu_{1i})} & 0 & 0 & 0 & 0 \\ 0 & \sqrt{\mu_{2i}(1-\mu_{2i})} & 0 & 0 & 0 \\ 0 & 0 & \sqrt{\mu_{3i}(1-\mu_{3i})} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\mu_{4i}(1-\mu_{4i})} & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\mu_{5i}(1-\mu_{5i})} \end{bmatrix}$$

$$\times \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \rho_{14}\sigma_1\sigma_4 & \rho_{15}\sigma_1\sigma_5 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \rho_{24}\sigma_2\sigma_4 & \rho_{25}\sigma_2\sigma_5 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 & \rho_{35}\sigma_3\sigma_5 \\ \rho_{14}\sigma_1\sigma_4 & \rho_{24}\sigma_2\sigma_4 & \rho_{34}\sigma_3\sigma_4 & \sigma_4^2 & \rho_{45}\sigma_4\sigma_5 \\ \rho_{15}\sigma_1\sigma_5 & \rho_{25}\sigma_2\sigma_5 & \rho_{35}\sigma_3\sigma_5 & \rho_{45}\sigma_4\sigma_5 & \sigma_5^2 \end{bmatrix}$$

$$\times \begin{bmatrix} \sqrt{\mu_{1i}(1-\mu_{1i})} & 0 & 0 & 0 & 0 \\ 0 & \sqrt{\mu_{2i}(1-\mu_{2i})} & 0 & 0 & 0 \\ 0 & 0 & \sqrt{\mu_{3i}(1-\mu_{3i})} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\mu_{4i}(1-\mu_{4i})} & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\mu_{5i}(1-\mu_{5i})} \end{bmatrix},$$

onde $\sqrt{\mu_{ri}(1-\mu_{ri})}$ denota a função de variância da distribuição binomial associada à r-ésima resposta $r = 1, \ldots 5$. Denote $\mathbf{R}$ a matriz de correlação entre as respostas, obtida pela multiplicação de cada elemento da matriz $\boldsymbol{\Omega}(\boldsymbol{\lambda})$ por $\frac{1}{\sigma_r \sigma_{r'}}$. Assim, para este particular conjunto de dados, a representação da matriz $\mathbf{R}$ é dada por:

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} & \rho_{25} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} & \rho_{35} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 & \rho_{45} \\ \rho_{15} & \rho_{25} & \rho_{35} & \rho_{45} & 1 \end{bmatrix},$$

onde $\rho_{rr'}$ denota a correlação entre a resposta r e r'.

Note que a matriz $\mathbf{R}$ é simétrica, não depende da estrutura da média e a quantidade de coeficientes de correlação desta matriz pode ser obtida por $R(R-1)/2$. Desse modo, esta matriz possui dez coeficientes de correlação. Assim, por exemplo, a correlação entre o percentual de gordura na região dos braços e pernas é denotada por ρ_{12}. Da mesma forma, as correlações entre o percentual de gordura na região dos braços e tronco, entre braços e androide e entre braços e ginecoide são representadas, respectivamente, por ρ_{13}, ρ_{14} e ρ_{15}. Reciprocamente, os coeficientes ρ_{23}, ρ_{24} e ρ_{25} indicam as correlações entre o percentual de gordura na região das pernas e tronco, entre pernas e androide e entre pernas e ginecoide. Já a correlação entre o percentual de gordura na região do tronco e androide é denotada por ρ_{34}, e entre tronco e ginecoide por ρ_{35}. Por fim, a correlação entre o percentual de gordura na região androide e ginecoide é expressa por ρ_{45}.

Numa primeira etapa da análise, ajustou-se o modelo de regressão multivariado e o univariado, no qual as análises foram feitas separadamente para cada variável resposta. O objetivo foi comparar as duas análises por meio de medidas de bondade de ajuste (*goodness-of-fit*). Dessa forma, a Tabela 6 apresenta o valor maximizado do

logaritmo da função de pseudo verossimilhança (plogLik), graus de liberdade (df), e valores dos pseudo critérios de informação de *Akaike* (pAIC) e Bayesiano (pBIC) para os modelos uni e multivariado.

TABELA 6 – VALOR MAXIMIZADO DO LOGARITMO DA FUNÇÃO DE PSEUDO VEROSSIMILHANÇA (plogLik), GRAUS DE LIBERDADE (df) E PSEUDO CRITÉRIOS DE INFORMAÇÃO DE AKAIKE (pAIC) E BAYESIANO (pBIC) PARA OS MODELOS UNI E MULTIVARIADO

Modelo	plogLik	df	pAIC	pBIC
Univariado	2176,19	35	−4282,38	−4096,65
Multivariado	3162,10	45	−6234,20	−5995,41

FONTE: O autor (2018).

De acordo com os resultados apresentados na Tabela 6, todas as medidas de bondade de ajuste apontam para o modelo multivariado como aquele que apresenta o melhor ajuste aos dados. A diferença entre os modelos uni e multivariado em termos de pseudo log-verossimilhança foi de 985,91, enquanto para os pseudo critérios de informação de *Akaike* e Bayesiano esta diferença foi de $-1951,82$ e $-1898,76$, respectivamente. Como o modelo multivariado apresentou vantagens na análise dos dados, todas as análises apresentadas a seguir são baseadas neste modelo.

Os resultados apresentados na Tabela 7 mostram que todas as covariáveis pertencentes aos preditores lineares (Equação 5.8) são significativamente diferentes de zero (p-valor $< 0,00$), com exceção da covariável idade para as variáveis respostas porcentagem de gordura na região das pernas (p-valor $= 0,90$) e ginecoide (p-valor $= 0,26$), respectivamente.

TABELA 7 – ESTATÍSTICA DE WALD (W_s), GRAUS DE LIBERDADE (df) E P-VALORES PARA OS COMPONENTES DO PREDITOR LINEAR DE CADA VARIÁVEL RESPOSTA

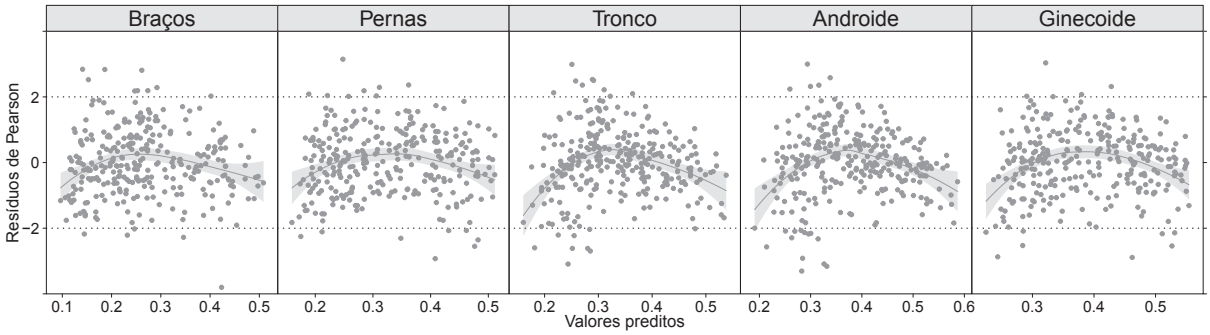
Efeitos	df	Braços %		Pernas %		Tronco %		Androide %		Ginecoide %	
		W_s	p-valor	W_s	p-valor	W_s	p-valor	W_s	p-valor	W_s	p-valor
IMC	1	201,97	$< 0,01$	122,66	$< 0,01$	337,07	$< 0,01$	253,47	$< 0,01$	136,35	$< 0,01$
Sexo	1	672,28	$< 0,01$	661,41	$< 0,01$	184,78	$< 0,01$	102,44	$< 0,01$	657,38	$< 0,01$
Idade	1	15,75	$< 0,01$	0,01	0,90	16,81	$< 0,01$	18,73	$< 0,01$	1,23	0,26
IPAQ	2	31,13	$< 0,01$	12,61	$< 0,01$	18,75	$< 0,01$	14,36	$< 0,01$	15,23	$< 0,01$

FONTE: O autor (2018).

Antes de interpretar os coeficientes estimados pelo modelo de regressão quase-beta multivariado, verificou-se a qualidade produzida pelo ajuste do modelo, por meio de uma análise de resíduos e de diagnóstico. Dessa forma, a Figura 24 apresenta os resíduos de Pearson versus os valores preditos pelo modelo para cada variável resposta.

A Figura 24 também apresenta curvas de suavização com bandas de confiança estimadas pelo método *loess* (CLEVELAND, 1979).

FIGURA 24 – RESÍDUOS DE PEARSON ASSOCIADOS AO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO PERCENTUAL DE GORDURA CORPORAL



FONTE: O autor (2018).

Os resíduos de Pearson mostrados na Figura 24 indicam que o modelo de regressão quase-beta multivariado ajustou-se satisfatoriamente aos dados do percentual de gordura corporal. Em geral, os resíduos variam entre -2 e 2 e observa-se que apenas alguns valores estão fora desse intervalo, o que não compromete os resultados da análise. Porém, vale ressaltar que algumas curvas de suavização se destacaram pela sua forma "quadrática", em especial, as curvas associadas aos percentuais de gordura na região do tronco, androide e ginecoide.

Nesse sentido, ajustou-se o modelo novamente incluindo o efeito de polinômios do 2º grau nas covariáveis contínuas (idade e IMC), além de verificar o efeito de interação entre as covariáveis do estudo. Após essa nova análise, verificou-se que nenhuma dessas possibilidades trouxe benefícios no ajuste do modelo, assim como na melhoria dos resíduos apresentados na Figura 24. Acredita-se que uma das principais causas do efeito "quadrático" observado nas curvas de suavização seja devido a falta de covariáveis não observadas no estudo.

Diante disso, foi proposta uma transformação para as variáveis respostas, afim de investigar o comportamento dos resíduos. Tal transformação, é simplesmente um reescalonamento das variáveis respostas (Y_{ri}) para o intervalo $[0,1]$. Note que, no reescalonamento, as variáveis respostas podem assumir os valores zero ou um, o que não acontece na escala original. Logo, a transformação proposta é dada por:

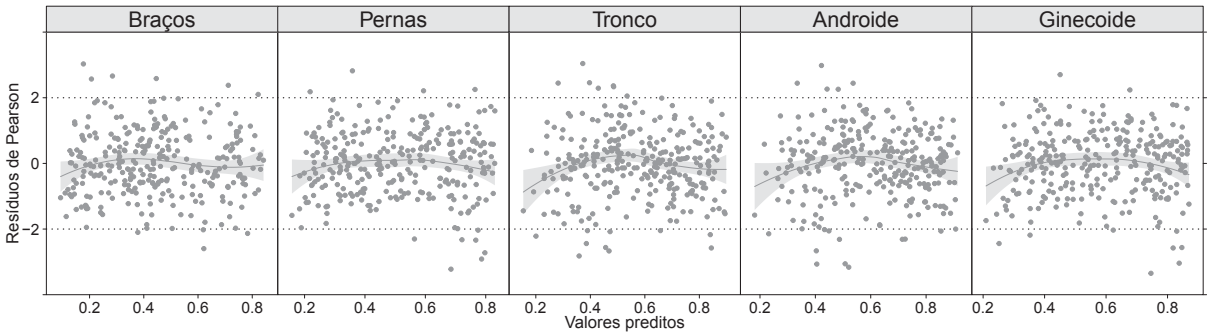
$$Y_{ri}^{*} = \frac{Y_{ri} - \min(Y_r)}{\max(Y_r) - \min(Y_r)}, \tag{5.9}$$

onde $\min(Y_r)$ e $\max(Y_r)$ denotam os valores mínimo e máximo da r-ésima resposta, para $r = 1,\ldots,5$. Para ilustar essa transformação, considere Y_{1i} o percentual de gor-

dura na região dos braços do i-ésimo indivíduo. Seus valores mínimo e máximo são dados, respectivamente, por $0,042$ e $0,547$. Assim, para o primeiro indivíduo da amostra, Y_{11}, observou-se um percentual de gordura nos braços igual a $0,163$ e por meio da Equação 5.9, esse valor em escala transformada fica $Y_{11}^{*} = 0,239$. Para as demais variáveis respostas, o cálculo é feito de forma semelhante. No Apêndice E, encontra-se uma tabela com as seis primeiras linhas do conjunto de dados, mostrando as variáveis respostas nas escalas original e transformada.

A Figura 25 apresenta os resíduos de Pearson versus os valores peditos pelo modelo, com as variáveis respostas em escala transformada (Equação 5.9). Dessa forma, os resultados apresentados na Figura 25, mostram que o efeito "quadrático" nas curvas de suavização diminuíram, melhorando de forma geral os resíduos do modelo.

FIGURA 25 – RESÍDUOS DE PEARSON ASSOCIADOS AO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO PERCENTUAL DE GORDURA CORPORAL EM ESCALA TRANSFORMADA (Y_r^{*})



FONTE: O autor (2018).

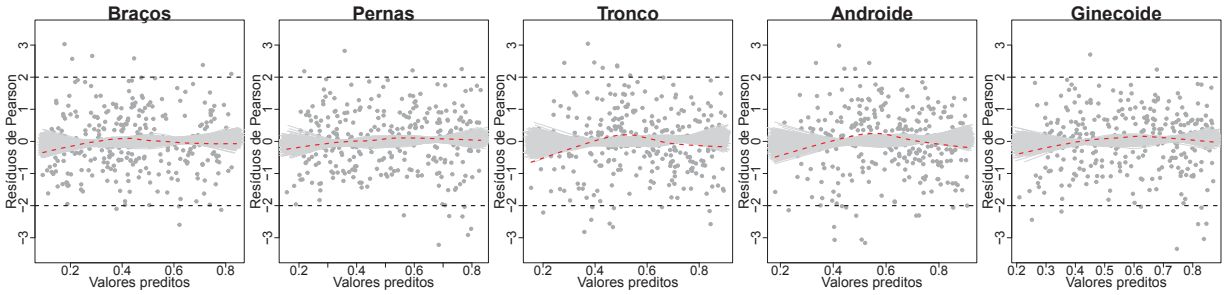
Com objetivo de construir bandas de confiança para os resíduos em escala transformada, adotou-se o seguinte procedimento. Considere $\mathbf{Y}_i^S = (Y_{1i}^S, \ldots, Y_{5i}^S)^\top$ um vetor de variáveis respostas simulado associado, respectivamente, aos percentuais de gordura nos braços, pernas, tronco, andróide e ginecóide do indivíduo i. Nesse caso, as variáveis respostas Y_{ri}^S foram simuladas a partir dos parâmetros estimados pelo modelo de regressão quase-beta multivariado com a escala transformada. Assim, a média de cada resposta simulada, foi obtida pelos valores preditos pelo modelo, isto é, $\mu_{ri}^S = \hat{\mu}_{ri}$. Os parâmetros de dispersão (σ_r^2) também foram transformados, para que fiquem na mesma escala do parâmetro de dispersão da densidade beta (Equação 3.2). A transformação é dada por:

$$\phi_r^S = \frac{1 - \sigma_r^2}{\sigma_r^2},\ r = 1, \ldots, 5.$$

Note que, o inverso dessa transformação corresponde a parametrização adotada no modelo de regressão quase-beta multivariado para os parâmetros de dispersão, ou seja,

$\sigma_r^2 = 1/(1+\phi_r)$. Assim, as variáveis respostas Y_{ri}^S foram simuladas a partir da densidade beta (Equação 3.2) com os parâmetros μ_{ri}^S e ϕ_r^S, isto é, $Y_{ri}^S \sim \mathcal{B}(\mu_{ri}^S, \phi_r^S)$. Após esses procedimentos, o modelo de regressão quase-beta multivariado foi ajustado novamente, usando as variáveis respostas simuladas, e os resíduos de Pearson versus os valores preditos pelo modelo foram plotados na Figura 26 por meio de curvas de suavização usando o método *loess*. Esse procedimento foi repetido 1000 vezes, gerando as bandas de confiança mostradas na Figura 26. As linhas pontilhadas em vermelho mostram o ajuste do modelo em escala transformada Y_{ri}^*, enquanto as bandas de confiança (linhas em cinza) ilustram os resultados do procedimento supracitado.

FIGURA 26 – RESÍDUOS DE PEARSON E BANDAS DE CONFIANÇA ASSOCIADAS AO AJUSTE DO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AOS DADOS DO PERCENTUAL DE GORDURA CORPORAL EM ESCALA TRANSFORMADA (Y_r^*)

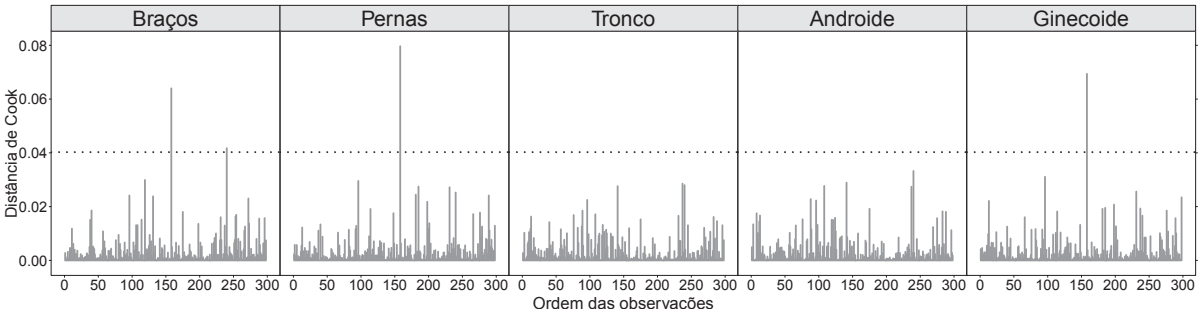


FONTE: O autor (2018).

Pelos resultados apresentados na Figura 26, é possível verificar que o ajuste do modelo mantém-se dentro das bandas de confiança dos dados simulados. No entanto, observa-se que alguns pontos das variáveis respostas relacionadas ao percentual de gordura na região do tronco e androide afastaram-se das bandas de confiança. Após investigar o comportamento dos resíduos em escala transformada, considere os resultados obtidos a seguir pelo modelo de regressão na escala original dos dados.

Assim, a Figura 27 mostra os gráficos da distância de Cook versus a ordem das observações para as cinco variáveis respostas. A linha pontilhada no centro da figura se refere ao ponto de corte $2p/n$. Portanto, valores acima de $0{,}0403$ indicam possíveis pontos influentes. Desta figura, observa-se que para os pecentuais de gordura na região dos braços, pernas e ginecoide existe apenas um ponto influente.

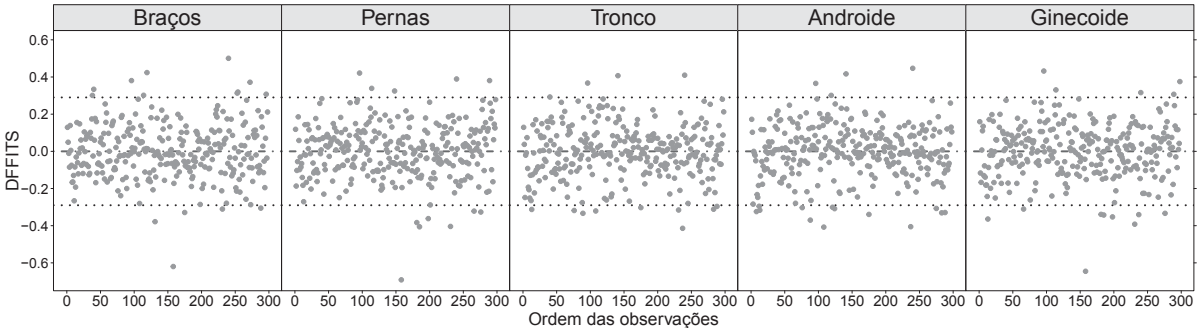
FIGURA 27 – DISTÂNCIA DE COOK PARA O MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO PERCENTUAL DE GORDURA CORPORAL



FONTE: O autor (2018).

Os resultados na Figura 28 mostram a medida DFFITS para cada variável resposta referente ao percentual de gordura corporal. A maioria dos pontos encontram-se no intervalo $\pm 0,284$, porém, observa-se alguns pontos abaixo do limite inferior.

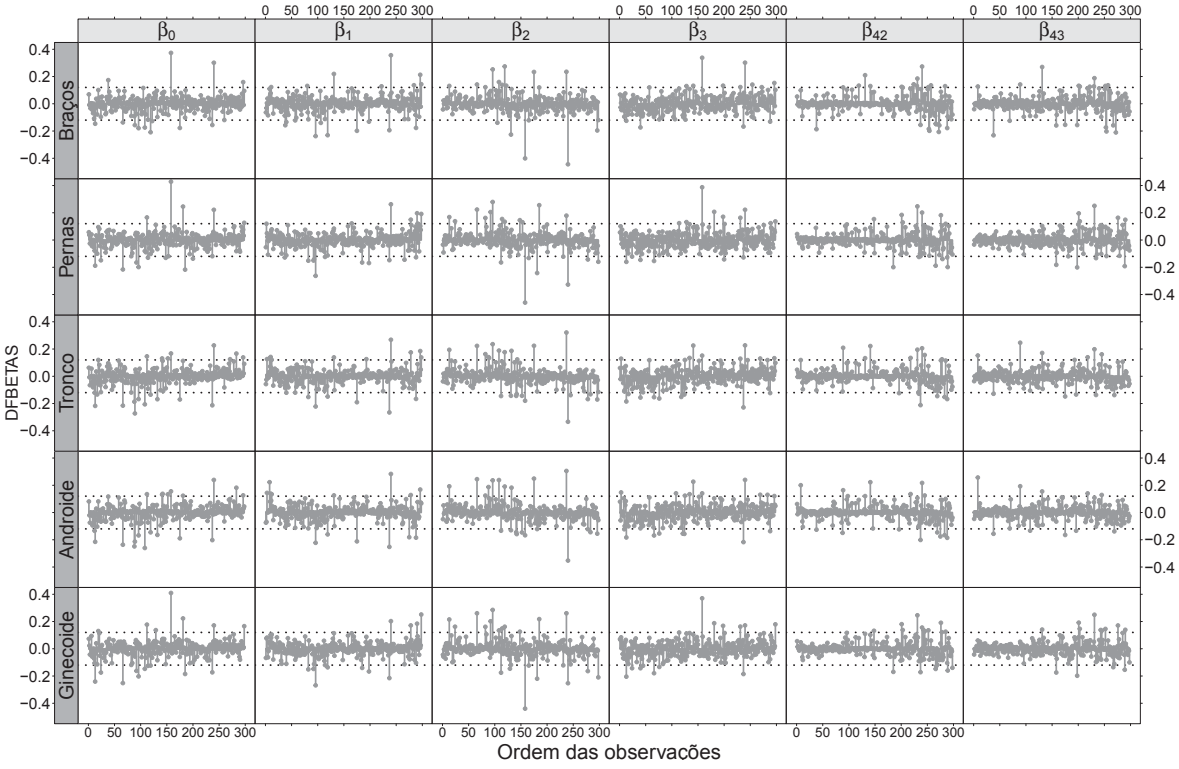
FIGURA 28 – DFFITS PARA O MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO PERCENTUAL DE GORDURA CORPORAL



FONTE: O autor (2018).

A Figura 29 apresenta a medida DFBETA para cada variável resposta. Por meio dessa medida, pode-se analisar a influência de cada observação sobre os coeficientes de regressão. Desse modo, valores fora do intervalo $\pm 0,116$ são considerados pontos influentes. De acordo com os resultados mostrados na Figura 29, existem pontos influentes no ajuste do modelo de regressão quase-beta multivariado. Os pontos que estão fora do intervalo, diferem entre as variáveis respostas. Além disso, a maior parte dos pontos encontram-se dentro do intervalo, o que mostra um bom ajuste do modelo aos dados.

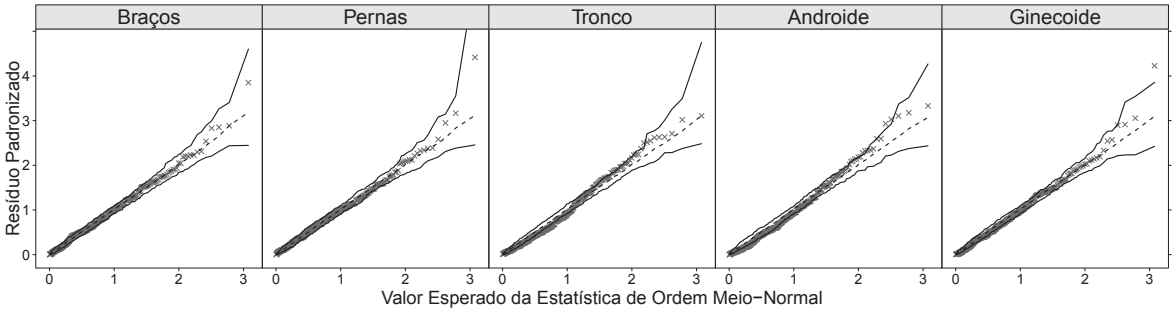
FIGURA 29 – DFBETAS PARA O MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO PERCENTUAL DE GORDURA CORPORAL



FONTE: O autor (2018).

Na Figura 30 encontram-se os gráficos de probabilidade meio-normal com envelope simulado. Os resultados apresentados nesta figura mostram que, em geral, o ajuste do modelo de regressão quase-beta multivariado está adequado para as cinco variáveis respostas, uma vez que os pontos encontram-se dentro das bandas de confiança. No entanto, para os percentuais de gordura nas regiões do tronco e androide pode-se observar alguns pontos fora do intervalo.

FIGURA 30 – GRÁFICO DE PROBABILIDADE MEIO-NORMAL COM ENVELOPE SIMULADO PARA CADA VARIÁVEL RESPOSTA DO PERCENTUAL DE GORDURA CORPORAL AJUSTADO PELO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO



FONTE: O autor (2018).

Com o intuito de investigar os pontos influentes e *outliers* que se destacaram na análise de diagnóstico, fez-se uma investigação no conjunto de dados examinando os pontos indicados na análise. Assim, conforme nos mostra a Figura 27, Figura 28 e Figura 29 o indivíduo 158 foi quem se destacou. Tal indivíduo é uma mulher, tem 49 anos, apresenta um IMC igual a 29,3 e tem percentuais de gordura nas regiões dos braços, pernas, tronco, androide e ginecoide iguais a: 0,196; 0,203; 0,378; 0,411 e 0,292 respectivamente. Em outras palavras, este indivíduo apresenta um elevado IMC, mas baixos valores para os percentuais de gordura na região dos braços, pernas e ginecoide conforme foi apontado na análise de diagnóstico. Dessa forma, mostrou-se que as técnicas de diagnóstico adaptadas para o modelo proposto são eficientes na detecção de pontos influentes e *outliers*.

De maneira geral, todas as medidas de diagnóstico além da análise de resíduos indicam que o modelo de regressão quase-beta multivariado apresentou um ajuste satisfatório aos dados do percentual de gordura corporal.

Após a análise de resíduos e de diagnóstico, serão apresentadas as principais interpretações relacionadas aos parâmetros de regressão estimados pelo modelo de regressão quase-beta multivariado, na escala original. Assim, a Tabela 8 apresenta as estimativas dos parâmetros e erros-padrão associados a cada variável resposta.

TABELA 8 – ESTIMATIVAS DOS PARÂMETROS (Est.) E ERROS-PADRÃO (EP) PARA O PERCENTUAL DE GORDURA NA REGIÃO DOS BRAÇOS, PERNAS, TRONCO, ANDROIDE E GINECOIDE, RESPECTIVAMENTE, OBTIDOS PELO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO

	Braços	Pernas	Tronco	Androide	Ginecoide
Efeito	Est.(EP)	Est.(EP)	Est.(EP)	Est.(EP)	Est.(EP)
`Intercepto`	-2,810(0,136)	-1,879(0,131)	-2,998(0,121)	-2,859(0,139)	-1,442(0,113)
`Idade`	0,004(0,001)	0,001(0,010)*	0,003(0,001)	0,004(0,001)	-0,009(0,008)*
`IMC`	0,086(0,006)	0,065(0,005)	0,098(0,005)	0,098(0,006)	0,059(0,005)
`Sexo-M`	-0,889(0,034)	-0,852(0,033)	-0,401(0,029)	-0,343(0,033)	-0,726(0,028)
`IPAQ-IA`	-0,118(0,047)	-0,044(0,046)	-0,106(0,042)	-0,115(0,048)	-0,066(0,041)
`IPAQ-A`	-0,248(0,045)	-0,145(0,044)	-0,175(0,040)	-0,177(0,046)	-0,146(0,038)
$\hat{\sigma}_r^2$	0,014(0,001)	0,015(0,001)	0,013(0,001)	0,018(0,001)	0,012(0,001)

Nota: *Indica p-valor $> 0,05$; `Sexo-M` = sexo (masculino); `IPAQ-IA` = IPAQ (insuficientemente ativo); `IPAQ-A` = IPAQ (ativo); $\hat{\sigma}_r^2$: coeficiente de dispersão associado à r-ésima resposta, $r = 1,\ldots,5$.

Com base nos resultados apresentados na Tabela 8, pode-se observar que a idade foi a única covariável que não se relacionou com todas as respostas. Assim, não faz diferença a idade dos indivíduos na avaliação do percentual de gordura na região das pernas e ginecoide. Entretanto, para as outras variáveis respostas, essa covariável mostrou-se relevante, podendo-se observar sinais positivos em todos os coeficientes.

Desse modo, quanto maior for a idade dos indivíduos maior se espera que sejam os percentuais de gordura na região dos braços, tronco e androide. O IMC também exerceu influência positiva nos percentuais de gordura corporal, tanto para homens quanto para mulheres. Como já era esperado, quanto maior o IMC dos indivíduos maior serão seus percentuais de gordura corporal.

O nível de atividade física, estimado pelo questionário IPAQ, mostrou ser um fator determinante na redução do percentual gordura corporal. Logo, há um indicativo de que os índivíduos classificados pelo IPAQ em ativos apresentam menos gordura distribuída pelo corpo do que os indivíduos classificados em sedentários, sendo esta avaliação para ambos os sexos. De maneira análoga, os indivíduos classificados em insuficientemente ativos também apresentam menos gordura corporal quando comparados aos sedentários. Porém, a redução de gordura corporal é maior para os indivíduos ativos. Basta olhar os valores dos coeficientes de regressão associados a esta covariável (`IPAQ-A`). Eles são maiores do que os coeficientes que medem o efeito do nível de atividade física dos indivíduos classificados em insuficientemente ativos (`IPAQ-IA`).

No ajuste do modelo de regressão quase-beta multivariado, fez-se uso da função de ligação *logit* que permite interpretar os coeficientes de regressão em termos de razão de chances. Nesse sentido, o Apêndice E apresenta uma tabela com razões de chances e intervalos com 95% de confiança para as estimativas dos parâmetros apresentadas na Tabela 8.

A correlação entre as variáveis respostas, dada a presença das covariáveis no modelo é apresentada na matriz a seguir:

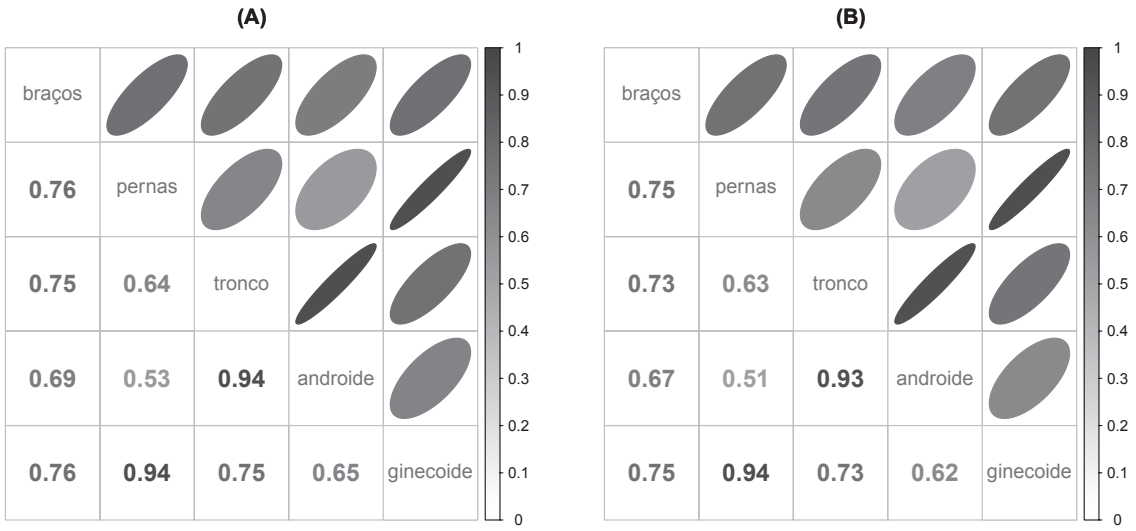
$$\hat{\mathbf{R}} = \begin{bmatrix} 1 & & & & \\ 0{,}7636(0{,}0359) & 1 & & & \\ 0{,}7466(0{,}0337) & 0{,}6442(0{,}0411) & 1 & & \\ 0{,}6938(0{,}0381) & 0{,}5343(0{,}0465) & 0{,}9433(0{,}0337) & 1 & \\ 0{,}7649(0{,}0353) & 0{,}9417(0{,}0379) & 0{,}7529(0{,}0363) & 0{,}6521(0{,}0418) & 1 \end{bmatrix}, \quad (5.10)$$

onde os números dentro dos parênteses denotam os erros padrões associados aos coeficientes de correlação. Todas as correlações apresentadas na matriz 5.10 são significativamente diferentes de zero. Assim, por exemplo, a correlação entre o percentual de gordura na região dos braços e pernas foi estimada em $\hat{\rho}_{12} = 0{,}7636(0{,}0359)$. Já a correlação entre o percentual de gordura na região das pernas e tronco foi estimada em $\hat{\rho}_{23} = 0{,}6442(0{,}0411)$. A maior correlação estimada foi entre o percentual de gordura na região do tronco e androide, $\hat{\rho}_{34} = 0{,}9433(0{,}0337)$.

A Figura 31 mostra as correlações entre os percentuais de gordura corporal estimadas pelos modelos de regressão quase-beta multivariado nas escalas original e transformada. Com base nos resultados apresentados na Figura 31, é possível observar

que as correlações entre os dois ajustes ficaram muito próximas uma da outra. Tal resultado, mostrou que as correlações não foram afetadas pelas transformações das variáveis respostas.

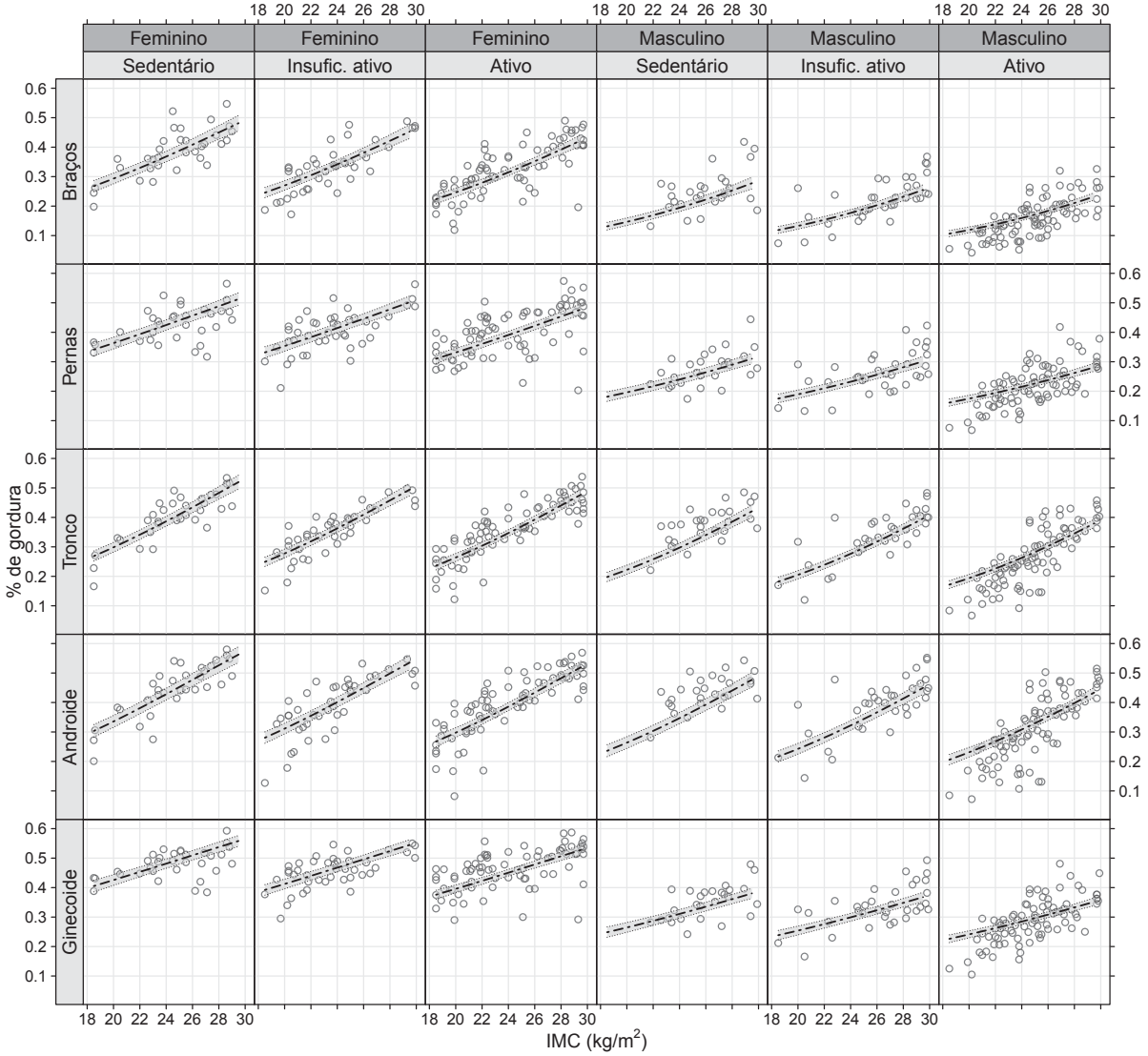
FIGURA 31 – IMAGEM DA MATRIZ DE CORRELAÇÃO ENTRE AS VARIÁVEIS RESPOSTAS ESTIMADA PELO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO NAS ESCALAS ORIGINAL (A) E TRANSFORMADA (B)



FONTE: O autor (2018).

Para tornar mais visual as variações entre os percentuais de gordura corporal conforme as covariáveis em estudo, construiu-se a Figura 32. Esta figura foi construída com base nos coeficientes de regressão e seus respectivos erros-padrão estimados pelo modelo de regressão quase-beta multivariado (Tabela 8). Assim, a Figura 32 apresenta curvas de predição com bandas de confiança (95%) para a média das variáveis respostas em função das covariáveis do estudo. A idade foi a única covariável que não se relacionou com todas as respostas. Logo, a Figura 32 foi construída fixando-se a idade média $(\bar{x} = 46$ anos$)$ dos indivíduos.

FIGURA 32 – GRÁFICO DO PERCENTUAL DE GORDURA ESTIMADO PARA CADA REGIÃO DO CORPO E INTERVALOS COM 95% DE CONFIANÇA OBTIDOS PELO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO EM FUNÇÃO DO IMC, SEXO (F-FEMININO OU M-MASCULINO) E IPAQ (SEDENTÁRIO, INSUFICIENTEMENTE ATIVO OU ATIVO)



FONTE: O autor (2018).

Com base nos resultados apresentados na Figura 32, pode-se observar que o percentual de gordura na região dos braços apresentou mudanças consideráveis com relação à prática de atividade física e o sexo dos indivíduos. Também observa-se uma variação semelhante com relação ao percentual de gordura na região das pernas. No entanto, essas variações não correspondem para os percentuais de gordura nas regiões do tronco e androide.

6 CONSIDERAÇÕES FINAIS

O objetivo geral desta dissertação foi propor um novo modelo de regressão para análise de variáveis respostas limitadas multivariada. O modelo foi especificado usando apenas suposições de primeiro e segundo momentos. Para estimação dos parâmetros, adotou-se uma abordagem que combina as funções de estimação quase-score e Pearson para estimação dos parâmetros de regressão e dispersão, respectivamente. Assim, o modelo proposto nesta dissertação segue o estilo de quase-verossimilhança apresentado por Wedderburn (1974), onde a especificação do modelo é feita pela combinação da função de variância da distribuição binomial com as tradicionais funções de ligação para dados binários.

A especificação baseada em momentos apresenta ao menos duas vantagens sobre os tradicionais métodos de estimação. Primeiro, não é preciso assumir uma distribuição de probabilidade multivariada para o vetor de variáveis respostas. Segundo, o algoritmo de estimação é de fácil implementação, podendo ser resumido a um simples e eficiente algoritmo do tipo Newton-score. Uma vantagem adicional da combinação entre as funções de estimação é a propriedade de insensitividade. Essa propriedade é análoga a ortogonalidade no contexto de estimação por máxima verossimilhança, o que também contribui para a simplificação do algoritmo de estimação (JØRGENSEN; KNUDSEN, 2004; BONAT; JØRGENSEN, 2016). Por outro lado, uma desvantagem da especificação baseada em momentos está na perda de eficiência dos estimadores. Consequentemente, os erros-padrões associados aos coeficientes de regressão são maiores, implicando em perda de poder dos testes de hipóteses.

No decorrer do trabalho foram apresentados três estudos de simulação. O primeiro foi delineado com objetivo de avaliar o comportamento do algoritmo NORTA na simulação de variáveis aleatórias beta correlacionadas. Nesse caso, o principal interesse foi identificar os valores mínimos e máximos que a matriz de correlação entre duas variáveis aleatórias beta pode assumir em função das suas médias marginais e valores dos parâmetros de dispersão. Os resultados desta avaliação mostraram que o espaço paramétrico da correlação é bastante reduzido quando se tem baixos valores para os parâmetros de dispersão associados com altos/baixos valores das médias marginais.

A construção e simulação de distribuições beta multivariadas é difícil devido ao fato que o suporte das distribuições marginais é o intervalo unitário. Tal restrição impõe um relacionamento entre a média e a variância da distribuição que por sua vez induz restrições não triviais nos possíveis valores das correlações entre os componentes do vetor aleatório. Além disso, tais restrições dependem da média e da variância da distribuição marginal especificada. Entender tais restrições foi fundamental para

avaliar o comportamento e aplicabilidade do algoritmo de simulação. Também houve a tentativa de simular variáveis aleatórias da distribuição de probabilidade simplex, que é uma alternativa à distribuição beta, porém os resultados não foram satisfatórios. Neste caso, o algoritmo NORTA exigiu um excessivo tempo computacional e não alcançou o objetivo proposto.

O segundo estudo de simulação visou avaliar as propriedades dos estimadores para os parâmetros de dispersão, no contexto de análise de dados longitudinais. Foram criados 36 cenários de simulação pela combinação de três estruturas de covariância, assumindo diferentes níveis de correlação, com quatro valores do parâmetro de dispersão da densidade beta. Os resultados desta avaliação mostraram que para pequenos valores de ϕ os estimadores são viciados. Porém, conforme aumentam-se os valores dos parâmetros de dispersão, bem como o tamanho da amostra o viés médio e o erro-padrão médio tendem a zero, mostrando que os estimadores são consistentes e não viciados.

O terceiro estudo de simulação teve por objetivo avaliar o comportamento dos estimadores em estudos com múltiplas respostas correlacionadas. Nesse caso, foram criados 35 cenários de simulação considerando-se duas variáveis aleatórias beta correlacionadas. Os resultados mostraram que em todos os cenários de simulação os estimadores são consistentes e não viciados, com exceção dos cenários onde $\phi = 0,00001$. Pelos resultados apresentados na Figura 12, os coeficientes β_{11} e β_{12} mostraram-se viciados para pequenas amostras. Contudo, à medida em que aumenta-se o tamanho da amostra, ambos viés e erro-padrão médio tendem a zero. O parâmetro de correlação ρ_{12} também mostrou-se viciado sob os cenários com $\phi = 0,00001$ (Figura 13), especialmente para correlações fixas em $\pm 0,75$. Assim, ficou fácil ver que quando $\phi \to 0$ a variância da beta tende para $\text{Var}(Y) = \mu(1-\mu)$. Neste caso, os dados gerados são próximos ou iguais a zeros e uns, uma vez que se tem a relação entre média e variância da distribuição Bernoulli.

Na sequência, analisou-se os conjuntos de dados apresentados no Capítulo 2. O primeiro corresponde ao índice de qualidade da água de reservatórios de usinas hidrelétricas. O principal desafio deste conjunto de dados está na análise de dados longitudinais e agrupados. Nesse sentido, modelou-se a estrutura de covariância do modelo proposto, afim de investigar possíveis correlações intraunidades amostrais. Dessa forma, foram propostas quatro estruturas de covariância, especificadas pela combinação de matrizes conhecidas. A comparação entre os modelos foi feita por meio de medidas de bondade de ajuste (`plogLik`, `pAIC` e `pBIC`). Os resultados mostraram que uma estrutura de covariância mais completa se ajustou melhor aos dados, permitindo estimar correlações intraunidades amostrais entre os locais e trimestres. Com base nos resultados obtidos, verificou-se a flexibilidade do modelo proposto para lidar com dados limitados em estudos longitudinais. Além disso, mostrou-se a importância da

modelagem da estrutura de covariância na análise de dados com as características acima mencionadas.

O segundo conjunto de dados se refere ao percentual de gordura corporal, que foi medido em cinco regiões do corpo e representam as variáveis respostas. Pelos métodos estatísticos convencionais, é difícil a análise de dados com tais características, principalmente, porque são dados multivariados limitados ao intervalo unitário. O principal objetivo dessa análise foi relacionar os percentuais de gordura corporal (braços, pernas, tronco, andróide e ginecóide) com a idade, sexo, nível de atividade física (IPAQ) e IMC de indivíduos avaliados no HC-UFPR. Os resultados mostraram que o modelo de regressão quase-beta multivariado apresentou um melhor desempenho na análise dos dados, quando comparado às análises feitas separadamente para cada variável resposta (modelo univariado). Tal comparação foi feita pelas medidas de bondade de ajuste (`plogLik`, `pAIC` e `pBIC`) apresentadas por Bonat (2018). Além disso, pelo modelo de regressão quase-beta multivariado foi possível estimar a matriz de correlação entre as respostas, dada a presença das covariáveis no modelo. De certa forma, tal abordagem trouxe mais informação na análise dos dados, indicando vantagens em usar o modelo proposto. Outro objetivo da dissertação foi adaptar técnicas de diagnóstico para o modelo proposto. Assim, no decorrer da análise dos dados, mostrou-se a aplicação e utilidade das medidas DFFITS, DFBETAS, distância de Cook e do gráfico de probabilidade meio-normal com envelope simulado para detecção de pontos influentes e *outliers*.

Portanto, o modelo de regressão proposto nesta dissertação permitiu lidar com dados limitados em estudos longitudinais, além de dados limitados em estudos com múltiplas respostas correlacionadas. Além disso, pode-se acomodar facilmente dados no intervalo $[0,1]$, incluindo zeros e uns.

6.1 FUTUROS TRABALHOS

Sugere-se como futuros trabalhos:

- Propor testes de hipóteses e de comparações múltiplas multivariados para investigar o efeito de covariáveis em estudos com múltiplas respostas, avaliando:
  1. Taxa de erros tipo I e tipo II
  2. Poder do teste
- Propor critérios de seleção para a estrutura de covariância, no contexto de análise de dados longitudinais;
- Modelar a estrutura de dispersão em função de covariáveis (modelos duplos);
- Analisar outros conjuntos de dados.

REFERÊNCIAS

ABBASI, T.; ABBASI, S. A. *Water quality indices*. Amsterdam: Elsevier, 2012. Citado na página 21.

AGÊNCIA NACIONAL DE ÁGUAS. 2018. Disponível em: ⟨http://pnqa.ana.gov.br/indicadores-indice-aguas.aspx⟩. Citado na página 106.

AITCHISON, J. The statistical analysis of compositional data. Chapman and Hall London, 1986. Citado na página 31.

ATKINSON, A. Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis. Clarendon Press, 1985. Citado na página 42.

BARNDORFF-NIELSEN, O. E.; JØRGENSEN, B. Some parametric models on the simplex. *Journal of Multivariate Analysis*, v. 39, n. 1, p. 106–116, 1991. Citado na página 29.

BAYER, F. M.; BAYER, D. M.; PUMI, G. Kumaraswamy autoregressive moving average models for double bounded environmental data. *Journal of Hydrology*, v. 555, p. 385–396, 2017. Citado na página 30.

BELSLEY, D. A.; KUH, E.; WELSCH, R. E. *Regression diagnostics*. New York: Wiley & Sons, 1980. Citado 2 vezes nas páginas 42 e 43.

BONAT, W. et al. Modelling the covariance structure in marginal multivariate count models: Hunting in bioko island. *Journal of Agricultural, Biological and Environmental Statistics*, p. 1–19, 2017. Citado 2 vezes nas páginas 18 e 39.

BONAT, W. H. *mcglm: Multivariate Covariance Generalized Linear Models*. Brazil, 2016. R package version 0.4.0. Citado na página 41.

BONAT, W. H. Modelling mixed types of outcomes in additive genetic models. *The International Journal of Biostatistics*, v. 13, n. 2, p. 1–16, 2017. Citado na página 46.

BONAT, W. H. Multiple response variables regression models in R: The mcglm package. *Journal of Statistical Software*, v. 84, n. 4, p. 1–30, 2018. Citado 3 vezes nas páginas 36, 37 e 83.

BONAT, W. H.; JØRGENSEN, B. Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, v. 65, n. 5, p. 649–675, 2016. Citado 4 vezes nas páginas 39, 40, 41 e 81.

BONAT, W. H. et al. Extended poisson-tweedie: Properties and regression models for count data. *Statistical Modelling*, v. 18, n. 1, p. 24–49, 2018. Citado na página 40.

BONAT, W. H. et al. Flexible quasi-beta regression models for continuous bounded data. *Statistical Modelling*, p. (published online), 2018. Citado 2 vezes nas páginas 18 e 30.

BONAT, W. H.; RIBEIRO JR, P. J.; ZEVIANI, W. M. Regression models with response on the unit interval: Specification, estimation and comparison. *Biometric Brazilian Journal*, v. 30, n. 4, p. 415–431, 2012. Citado na página 29.

BONAT, W. H.; RIBEIRO JR, P. J.; ZEVIANI, W. M. Likelihood analysis for a class of beta mixed models. *Journal of Applied Statistics*, v. 42, n. 2, p. 252–266, 2015. Citado na página 30.

CARIO, M. C.; NELSON, B. L. *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix*, 1997. Citado 3 vezes nas páginas 33, 46 e 52.

CEPEDA-CUERVO, E.; ACHCAR, J. A.; LOPERA, L. G. Bivariate beta regression models: joint modeling of the mean, dispersion and association parameters. *Journal of Applied Statistics*, v. 41, n. 3, p. 677–687, 2014. Citado na página 31.

CHEN, H. Initialization for norta: Generation of random vectors with specified marginals and correlations. *INFORMS Journal on Computing*, v. 13, n. 4, p. 312–331, 2001. Citado na página 46.

CLEVELAND, W. S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, v. 74, n. 368, p. 829–836, 1979. Citado 3 vezes nas páginas 26, 60 e 72.

COOK, R. D. Detection of influential observation in linear regression. *Technometrics*, v. 19, n. 1, p. 15–18, 1977. Citado 2 vezes nas páginas 42 e 43.

COPEL. 2018. Disponível em: ⟨http://www.copel.com/hpcopel/root/nivel2.jsp?endereco=%2Fhpcopel%2Froot%2Fpagcopel2.nsf%2F044b34faa7cc1143032570bd0059aa29%2F7e60b7740cdc206003257412005e4734⟩. Citado na página 22.

DEMIDENKO, E. *Mixed Models: Theory and Applications with R*. New Jersey: Wiley, 2013. Citado na página 39.

DIAS, R. A. P. *Simulação estocástica de variáveis aleatórias Poisson correlacionadas: aplicação ao controle populacional do percevejo (Euschistus heros Fabricius) da soja (Glycine max L.)*. Tese (Doutorado) — Universidade de São Paulo, 2014. Citado na página 46.

DIGGLE, P. et al. *Analysis of Longitudinal Data (Second edition)*. United Kingdom: Oxford University Press, 2002. 400 p. Citado na página 22.

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. New York: John Wiley & Sons, 2014. Citado na página 17.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, v. 31, n. 7, p. 799–815, 2004. Citado 5 vezes nas páginas 18, 29, 31, 36 e 47.

FIGUEROA-ZÚÑIGA, J. I.; ARELLANO-VALLE, R. B.; FERRARI, S. L. Mixed beta regression: A bayesian perspective. *Computational Statistics & Data Analysis*, v. 61, p. 137–147, 2013. Citado na página 30.

FITZMAURICE, G. M.; LAIRD, N. M.; WARE, J. H. *Applied Longitudinal Analysis (Second edition)*. New Jersey: John Wiley and Sons Inc., 2011. 740 p. Citado na página 22.

GHOSH, S.; HENDERSON, S. G. Behavior of the norta method for correlated random vector generation as the dimension increases. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, v. 13, n. 3, p. 276–294, 2003. Citado na página 46.

GODAMBE, V. P.; THOMPSON, M. Some aspects of the theory of estimating equations. *Journal of Statistical Planning and Inference*, v. 2, n. 1, p. 95–104, 1978. Citado na página 41.

GRÜN, B.; KOSMIDIS, I.; ZEILEIS, A. Extended beta regression in R: shaken, stirred, mixed, and partitioned. *Journal of Statistical Software*, v. 48, n. 11, p. 1–25, 5 2012. Citado na página 29.

GRUNWALD, G. K.; RAFTERY, A. E.; GUTTORP, P. Time series of continuous proportions. *Journal of the Royal Statistical Society, Series B*, v. 55, n. 1, p. 103–116, 1993. Citado na página 30.

HIJAZI, R. H.; JERNIGAN, R. W. Modelling compositional data using dirichlet regression models. *Journal of Applied Probability & Statistics*, v. 4, n. 1, p. 77–91, 2009. Citado na página 31.

HUNGER, M.; DÖRING, A.; HOLLE, R. Longitudinal beta regression models for analyzing health-related quality of life scores over time. *BMC medical research methodology*, v. 12, n. 1, p. 144, 2012. Citado na página 30.

JØRGENSEN, B.; KNUDSEN, S. J. Parameter orthogonality and bias adjustment for estimating functions. *Scandinavian Journal of Statistics*, v. 31, n. 1, p. 93–114, 2004. Citado 4 vezes nas páginas 39, 40, 41 e 81.

KENDLER, D. L. et al. The official positions of the international society for clinical densitometry: indications of use and reporting of dxa for body composition. *Journal of Clinical Densitometry*, v. 16, n. 4, p. 496–507, 2013. Citado na página 25.

KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of variates observed on $(0,1)$: percentages, proportions and fractions. *Statistical Modelling*, v. 3, n. 3, p. 193–213, 2003. Citado 2 vezes nas páginas 18 e 29.

LEMONTE, A. J.; BARRETO-SOUZA, W.; CORDEIRO, G. M. The exponentiated kumaraswamy distribution and its log-transform. *Braz. J. Probab. Stat.*, v. 27, n. 1, p. 31–53, 2013. Citado na página 29.

LEMONTE, A. J.; BAZÁN, J. L. New class of johnson sb distributions and its associated regression model for rates and proportions. *Biometrical Journal*, v. 58, n. 4, p. 727–746, 2016. Citado 2 vezes nas páginas 18 e 30.

LI, S. T.; HAMMOND, J. L. Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Transactions on Systems, Man, and Cybernetics*, n. 5, p. 557–561, 1975. Citado na página 46.

LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, v. 73, n. 1, p. 13–22, 1986. Citado 2 vezes nas páginas 17 e 39.

LIU, F.; EUGENIO, E. C. A review and comparison of bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. *Statistical methods in medical research*, v. 27, n. 4, p. 1024–1044, 2018. Citado na página 29.

LÓPEZ, F. O. A Bayesian approach to parameter estimation in simplex regression model: A comparison with beta regression. *Revista Colombiana de Estadística*, v. 36, n. 6, p. 1–21, 2013. Citado na página 29.

MASAROTTO, G.; VARIN, C. et al. Gaussian copula marginal regression. *Electronic Journal of Statistics*, v. 6, p. 1517–1549, 2012. Citado na página 30.

MATSUDO, S. et al. Questionário internacional de atividade física (ipaq): Estudo de validade e reprodutibilidade no brasil. *Revista Brasileira de Atividade Física & Saúde*, v. 6, n. 2, p. 5–18, 2001. Citado na página 25.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models (Second edition)*. London: Chapman & Hall, 1989. 532 p. Citado na página 17.

MCKENZIE, E. An autoregressive process for beta random variables. *Management Science*, v. 31, n. 8, p. 988–997, 1985. Citado na página 30.

MITNIK, P. A.; BAEK, S. The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. *Statistical Papers*, v. 54, n. 1, p. 177–192, 2013. Citado na página 30.

MIYASHIRO, E. S. *Modelos de regressão beta e simplex para análise de proporções*. Dissertação (Mestrado) — Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2008. Citado na página 29.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. New York: John Wiley & Sons, 2012. Citado 3 vezes nas páginas 17, 23 e 42.

MOUSA, A. M.; EL-SHEIKH, A. A.; ABDEL-FATTAH, M. A. A gamma regression for bounded continuous variables. *Advances and Applications in Statistics*, v. 49, n. 4, p. 305, 2016. Citado 2 vezes nas páginas 18 e 30.

MURTEIRA, J. M.; RAMALHO, J. J. Regression analysis of multivariate fractional data. *Econometric Reviews*, v. 35, n. 4, p. 515–552, 2016. Citado na página 31.

MYERS, R. et al. *Generalized Linear Models: With Applications in Engineering and the Sciences: Second Edition*. New Jersey: John Wiley and Sons Inc., 2010. 496 p. Citado na página 17.

NAHAS, M. V. *Atividade física, saúde e qualidade de vida: conceitos e sugestões para um estilo de vida ativo*. Londrina: Midiograf, 2001. Citado na página 25.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, Wiley for the Royal Statistical Society, v. 135, n. 3, p. 370–384, 1972. Citado na página 17.

NETER, J. et al. *Applied linear statistical models*. Chicago: Irwin, 1996. Citado na página 44.

NIAKI, S. T. A.; ABBASI, B. Generating correlation matrices for normal random vectors in norta algorithm using artificial neural networks. *Journal of Uncertain Systems*, v. 2, n. 3, p. 192–201, 2008. Citado na página 46.

PAOLINO, P. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, v. 9, n. 4, p. 325–346, 2001. Citado na página 29.

PETAK, S. et al. The official positions of the international society for clinical densitometry: body composition analysis reporting. *Journal of Clinical Densitometry*, v. 16, n. 4, p. 508–519, 2013. Citado na página 25.

PETTERLE, R. R. et al. Comparação e aplicação de modelos de regressão binária na retenção de capacetes de motociclistas. *Revista Brasileira de Biometria*, v. 35, n. 2, p. 266–282, 2017. Citado na página 36.

PETTERLE, R. R. et al. Análise da composição corporal via modelos de regressão beta. *Revista Brasileira de Biometria*, v. 36, n. 2, p. 336–359, 2018. Citado na página 26.

QIU, Z.; SONG, P. X.-K.; TAN, M. Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics*, v. 35, n. 4, p. 577–596, 2008. Citado na página 30.

R DEVELOPMENT CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Citado 5 vezes nas páginas 34, 41, 47, 49 e 52.

ROCHA, A. V.; CRIBARI-NETO, F. Beta autoregressive moving average models. *Test*, v. 18, n. 3, p. 529–545, 2008. Citado na página 30.

SANTOS, B. P. d. *Implementação em R de modelos de regressão binária com ligação paramétrica*. Dissertação (Mestrado) - Universidade de São Paulo, São Paulo - SP, 2013. Citado na página 36.

SILVA, C. da; MIGON, H.; CORREIA, L. Dynamic Bayesian beta models. *Computational Statistics & Data Analysis*, v. 55, n. 6, p. 2074–2089, 2011. Citado na página 30.

SILVA, G. d. S. F. d. et al. Avaliação do nível de atividade física de estudantes de graduação das áreas saúde/biológica. *Revista Brasileira de Medicina do Esporte*, v. 13, n. 1, p. 39–42, 2007. Citado na página 25.

SIMAS, A. B.; BARRETO-SOUZA, W.; ROCHA, A. V. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, v. 54, n. 2, p. 348–366, 2010. Citado na página 29.

SMITHSON, M.; VERKUILEN, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, v. 11, n. 1, p. 54–71, 2006. Citado na página 29.

SONATI, J. *Qualidade de Vida e Composição Corporal: Características do Envelhecimento Bem Sucedido. 84 p.* Tese (Doutorado) - Universidade Estadual de Campinas, Campinas, 2012. Citado na página 25.

SONG, P. X.-K.; QIU, Z.; TAN, M. Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal*, v. 46, n. 5, p. 540–553, 2004. Citado na página 30.

SONG, P. X.-K.; TAN, M. Marginal models for longitudinal continuous proportional data. *Biometrics*, v. 56, n. 2, p. 496–502, 2000. Citado na página 30.

SOUZA, D. F.; MOURA, F. A. Multivariate beta regression with application in small area estimation. *Journal of Official Statistics*, v. 32, n. 3, p. 747, 2016. Citado na página 31.

SU, P. *NORTARA: Generation of Multivariate Data with Arbitrary Marginals*, 2014. R package version 1.0.0. Citado 4 vezes nas páginas 34, 47, 49 e 52.

TAN, M.; QU, Y.; KUTNER, M. H. Model diagnostics for marginal regression analysis of correlated binary data. *Communications in Statistics-Simulation and Computation*, v. 26, n. 2, p. 539–558, 1997. Citado na página 45.

VENEZUELA, M. K.; BOTTER, D. A.; SANDOVAL, M. C. Diagnostic techniques in generalized estimating equations. *Journal of Statistical Computation and Simulation*, v. 77, n. 10, p. 879–888, 2007. Citado 3 vezes nas páginas 42, 43 e 45.

VERBEKE, G. et al. The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, v. 23, n. 1, p. 42–59, 2014. Citado na página 28.

VERKUILEN, J.; SMITHSON, M. Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, v. 37, n. 1, p. 82–113, 2012. Citado na página 30.

WEDDERBURN, R. W. M. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, v. 61, n. 3, p. 439–447, 1974. Citado 2 vezes nas páginas 39 e 81.

ZEGER, S. L.; LIANG, K.-Y.; ALBERT, P. S. Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, v. 44, p. 1049–1060, 1988. Citado 2 vezes nas páginas 17 e 39.

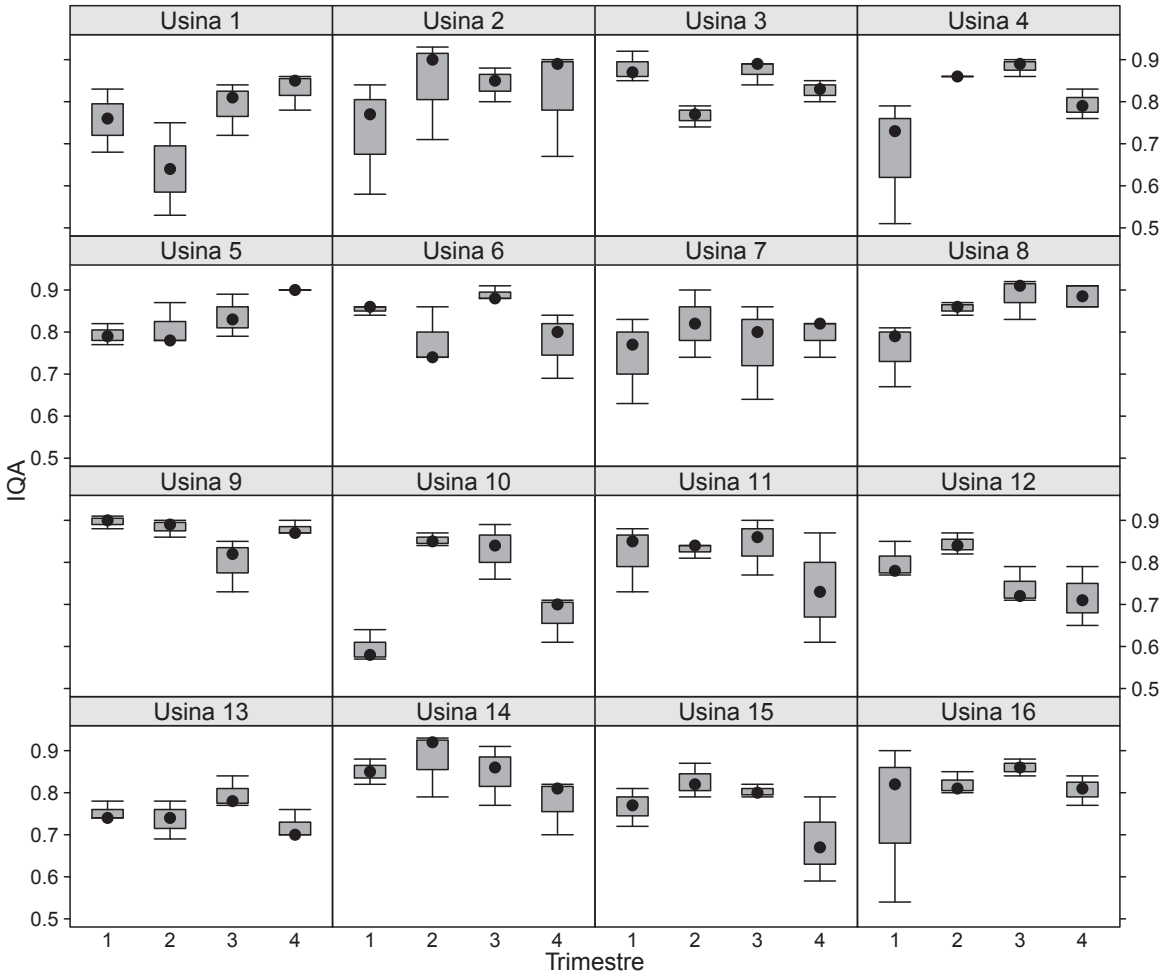
ZHAO, W.; LIAN, H.; BANDYOPADHYAY, D. A partially linear additive model for clustered proportion data. *Statistics in medicine*, v. 37, n. 6, p. 1009–1030, 2018. Citado na página 30.

ZHENG, X.; QIN, G.; TU, D. A generalized partially linear mean-covariance regression model for longitudinal proportional data, with applications to the analysis of quality of life data from cancer clinical trials. *Statistics in medicine*, v. 36, n. 12, p. 1884–1894, 2017. Citado na página 30.

Apêndices

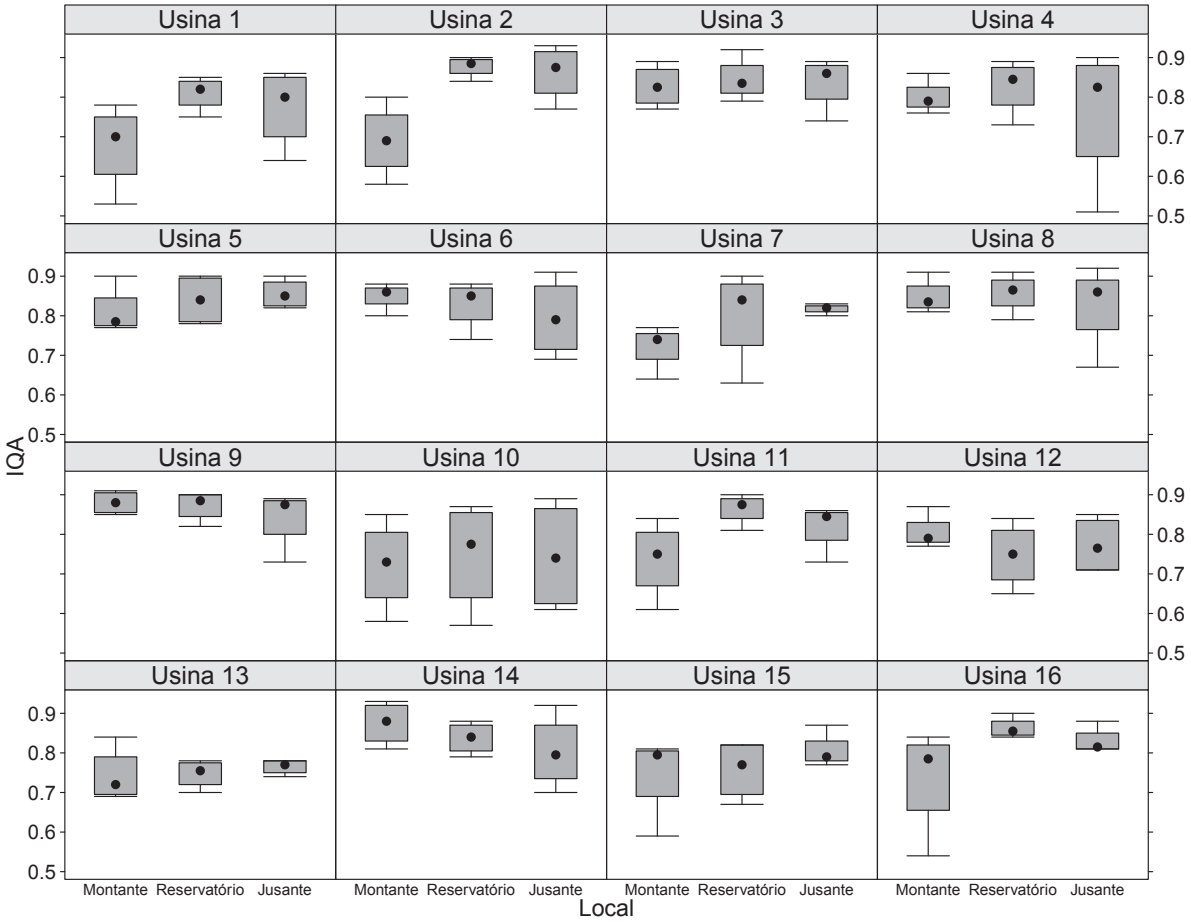
APÊNDICE A – GRÁFICOS BOXPLOTS PARA O CONJUNTO DE DADOS DO ÍNDICE DE QUALIDADE DA ÁGUA (IQA)

FIGURA 33 – GRÁFICOS BOXPLOTS PARA O IQA POR USINA E TRIMESTRE



FONTE: O autor (2018).

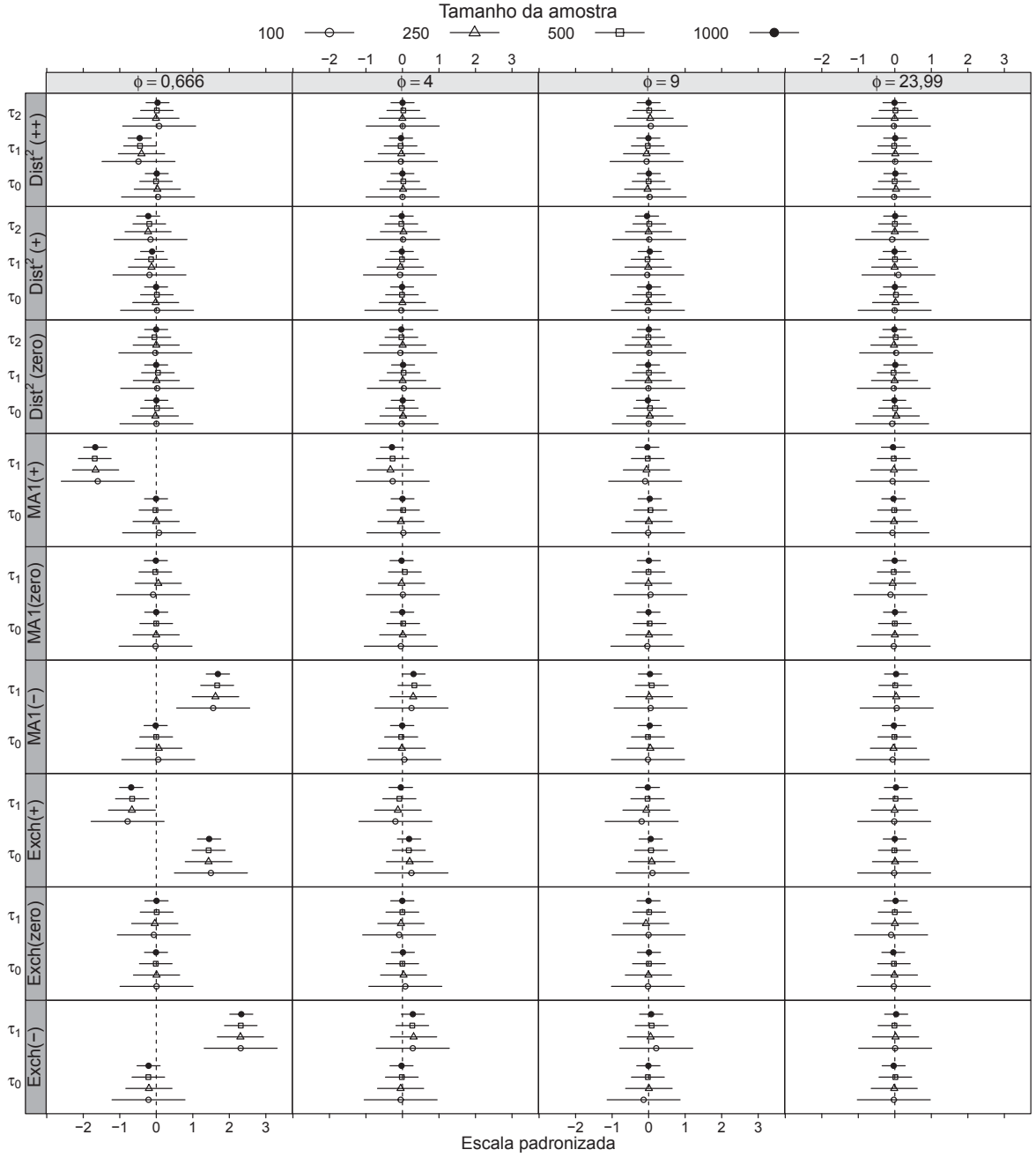
FIGURA 34 – GRÁFICOS BOXPLOTS PARA O IQA POR USINA E LOCAL



FONTE: O autor (2018).

APÊNDICE B – PROPRIEDADES DOS ESTIMADORES EM ESTUDOS LONGITUDINAIS

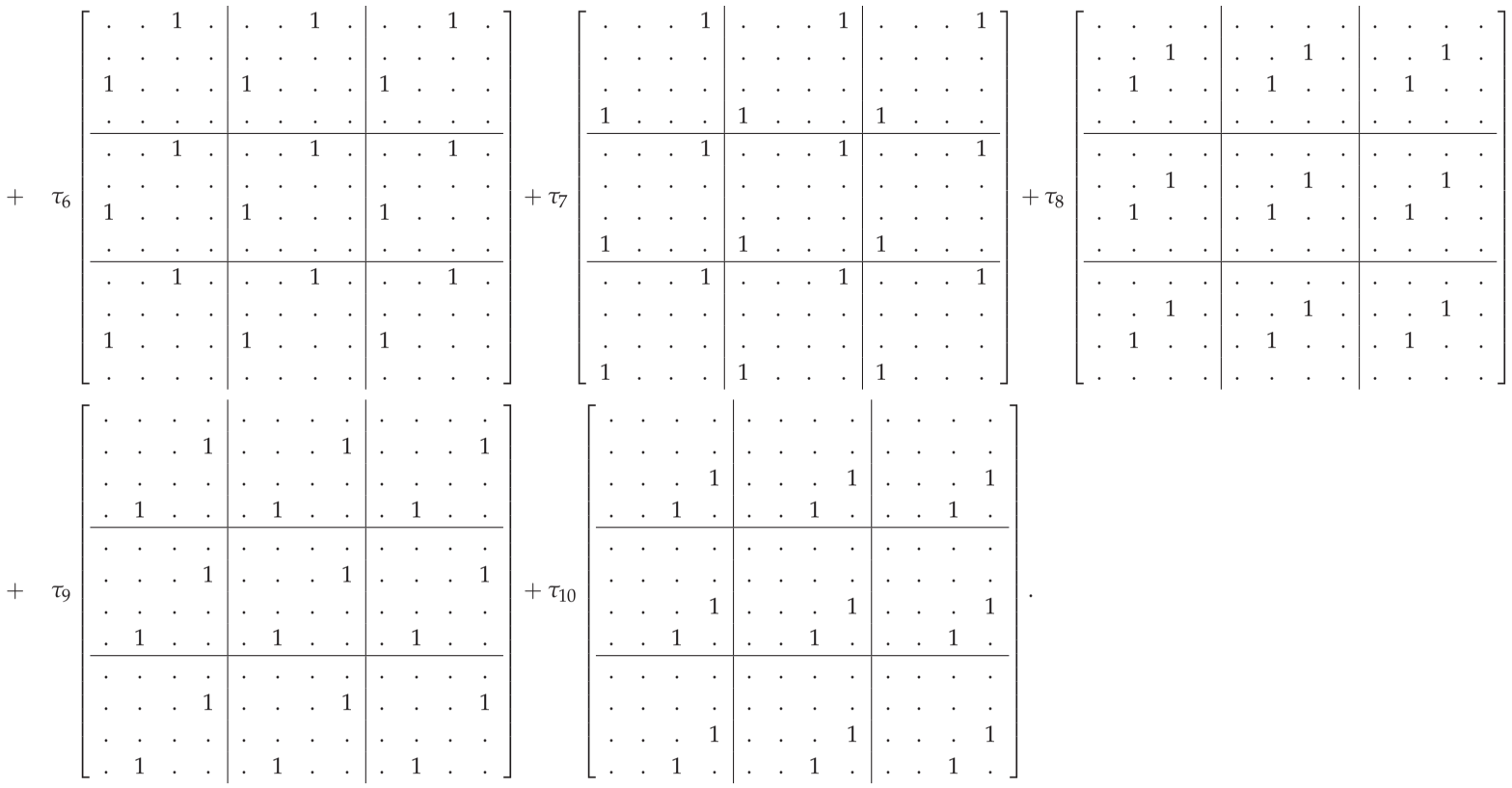
FIGURA 35 – VIÉS MÉDIO E INTERVALOS DE CONFIANÇA EM ESCALA PADRONIZADA PARA OS PARÂMETROS DE DISPERSÃO POR TAMANHO DE AMOSTRA E ESTRUTURA DE COVARIÂNCIA COM DIFERENTES NÍVEIS DE CORRELAÇÃO



FONTE: O autor (2018).

$$
\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix}
1 & . & . & . & . & . & . & . & . & . & . & . \\
. & 1 & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & . & . & . & . & . & . & . \\
. & . & . & 1 & . & . & . & . & . & . & . & . \\
. & . & . & . & 1 & . & . & . & . & . & . & . \\
. & . & . & . & . & 1 & . & . & . & . & . & . \\
. & . & . & . & . & . & 1 & . & . & . & . & . \\
. & . & . & . & . & . & . & 1 & . & . & . & . \\
. & . & . & . & . & . & . & . & 1 & . & . & . \\
. & . & . & . & . & . & . & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & 1 & . \\
. & . & . & . & . & . & . & . & . & . & . & 1
\end{bmatrix} + \tau_1 \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix} + \tau_2 \left[\begin{array}{cccc|cccc|cccc|cccc}
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
1 & . & . & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\
\hline
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
1 & . & . & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\
\hline
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
1 & . & . & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\
\hline
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
1 & . & . & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & . & .
\end{array}\right]
$$

$$
+ \tau_3 \left[\begin{array}{ccc|ccc|ccc|ccc}
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
\hline
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
\hline
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
\hline
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & .
\end{array}\right] + \tau_4 \left[\begin{array}{ccc|ccc|ccc|ccc}
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\
\hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\
\hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\
\hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & .
\end{array}\right] + \tau_5 \left[\begin{array}{ccc|ccc|ccc|ccc}
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\
1 & . & . & . & 1 & . & . & 1 & . & . & 1 & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
\hline
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\
1 & . & . & . & 1 & . & . & 1 & . & . & 1 & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
\hline
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\
1 & . & . & . & 1 & . & . & 1 & . & . & 1 & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & .
\end{array}\right]
$$


+ τ6
+ τ7
+ τ8
+ τ9
+ τ10

APÊNDICE D – CÓDIGOS EM R USADOS NA ANÁLISE DOS DADOS DO ÍNDICE DE QUALIDADE DA ÁGUA (IQA)

```
# Data analysis I: Water quality index
# Author: Ricardo Rasmussen Petterle/UFPR
# Loading extra packages
require(Matrix)
require(mcglm)
# Loading extra functions
source("mc_non_aux.R")
source("mc_odds_ratio.R")
# Loading data set
da <- read.table("IQADataSet.txt", h = T)
# Preparing data set
da$TRIM <- as.factor(da$TRIM)
da$LOCAL <- factor(da$LOCAL, levels = levels(da$LOCAL)[c(2,3,1)])
cod_na <- as.numeric(rownames(da[is.na(da$y),])) # remove NA's
data_na <- da[-cod_na,]
# Covariance structures
# Independent
Z0 <- mc_id(data_na)
# Exchangeable
Z1 <- mc_mixed(~ 0 + as.factor(id), data = data_na)
# Unstructured 1 (local)
Z2_ini <- mc_non_aux(3)
Z12 <- list()
Z13 <- list()
Z23 <- list()
II <- Matrix(rep(1, 16),4,4)
for(i in 1:16) {
  Z12[[i]] <- kronecker(II, Z2_ini[[1]])
  Z13[[i]] <- kronecker(II, Z2_ini[[2]])
  Z23[[i]] <- kronecker(II, Z2_ini[[3]])}
Z12 <- bdiag(Z12)
Z13 <- bdiag(Z13)
Z23 <- bdiag(Z23)
Z2 <- list(Z12, Z13, Z23)
Z2_na <- mc_remove_na(Z2, cod = cod_na)
```

```
# Unstructured 2 (quarter)
Z3_ini <- mc_non_aux(4)
II <- Matrix(rep(1,9), 3, 3)
Zt1t2 <- kronecker(Z3_ini[[1]], II)
Zt1t3 <- kronecker(Z3_ini[[2]], II)
Zt1t4 <- kronecker(Z3_ini[[3]], II)
Zt2t3 <- kronecker(Z3_ini[[4]], II)
Zt2t4 <- kronecker(Z3_ini[[5]], II)
Zt3t4 <- kronecker(Z3_ini[[6]], II)
Z3_12 <- list()
Z3_13 <- list()
Z3_14 <- list()
Z3_23 <- list()
Z3_24 <- list()
Z3_34 <- list()
for(i in 1:16) {
  Z3_12[[i]] <- Zt1t2
  Z3_13[[i]] <- Zt1t3
  Z3_14[[i]] <- Zt1t4
  Z3_23[[i]] <- Zt2t3
  Z3_24[[i]] <- Zt2t4
  Z3_34[[i]] <- Zt3t4
}
Z3 <- list(bdiag(Z3_12), bdiag(Z3_13),bdiag(Z3_14),
           bdiag(Z3_23), bdiag(Z3_24), bdiag(Z3_34))
Z3_na <- mc_remove_na(Z3, cod = cod_na)
# Linear predictor
form <- y ~ LOCAL + TRIM
data_na$Ntrial <- 1
# Fitting
# Independent
fit_id <- mcglm(linear_pred = c(form),
                variance = "binomialP",
                link = "logit",
                Ntrial = list(data_na$Ntrial),
                matrix_pred = list(c(Z0)),
                data = data_na,
                control_algorithm = list(tol = 1e-04,
                                         tuning = 1.0))
```

```
# Exchangeable
fit_cp <- mcglm(linear_pred = c(form),
                variance = "binomialP",
                link = "logit",
                Ntrial = list(data_na$Ntrial),
                matrix_pred = list(c(Z0,Z1)),
                data = data_na,
                control_algorithm = list(tol = 1e-04,
                                         tuning = 0.8))
# Unstructured 1
fit_un1 <- mcglm(linear_pred = c(form),
                 variance = "binomialP",
                 link = "logit",
                 Ntrial = list(data_na$Ntrial),
                 matrix_pred = list(c(Z0,Z1,Z2_na)),
                 data = data_na,
                 control_algorithm = list(tol = 1e-04,
                                          tuning = 0.8))
# Unstructured 2
fit_un2 <- mcglm(linear_pred = c(form),
                 variance = "binomialP",
                 link = "logit",
                 Ntrial = list(data_na$Ntrial),
                 matrix_pred = list(c(Z0,Z1,Z2_na,Z3_na)),
                 data = data_na,
                 control_algorithm = list(tol = 1e-04,
                                          tuning = 0.8))
# Comparing models
rbind(gof(fit_id), gof(fit_cp), gof(fit_un1), gof(fit_un2))
# ANOVA Wald test
anova(fit_un2)
# Regression parameters
summary(fit_un2, print = "Regression")
# Odds ratio and CI 95%
names = c("Trimeste 2","Trimeste 3","Trimeste 4",
          "Reservatorio","Jusante")
mc_odds_ratio(fit = fit_un2, names = names, response = 1)
# Dispersion parameters
summary(fit_un2, print = "Dispersion")
```

```
# Intraclass correlation (local)
tau <- coef(fit_un2, type= "tau")$Estimates
ZZ <- mc_non_aux(3)
ZZ1 <- Matrix(rep(1, 9),3,3)
ZZ0 <- Diagonal(3,1)
ICC_LOCAL <- cov2cor(ZZ0*tau[1] + ZZ1*tau[2] + ZZ[[1]]*tau[3] +
                      ZZ[[2]]*tau[4] + ZZ[[3]]*tau[5])
# Intraclass correlation (quarter)
tau <- coef(fit_un2, type= "tau")$Estimates
ZZ <- mc_non_aux(4)
ZZ1 <- Matrix(rep(1, 16),4,4)
ZZ0 <- Diagonal(4,1)
ICC_TRIM <- cov2cor(ZZ0*tau[1] + ZZ1*tau[2] +
                     ZZ[[1]]*tau[6] + ZZ[[2]]*tau[7] + ZZ[[3]]*tau[8] +
                     ZZ[[4]]*tau[9] + ZZ[[5]]*tau[10] + ZZ[[6]]*tau[11])
# END
```

APÊNDICE E – RESULTADOS COMPLEMENTARES PARA O CONJUNTO DE DADOS DO PERCENTUAL DE GORDURA CORPORAL

Neste apêndice, como forma ilustrativa, encontram-se as seis primeiras linhas do conjunto de dados relacionado ao percentual de gordura corporal. São apresentadas as cinco variáveis respostas nas escalas original Y_r e transformada Y_r^*. Tanto numa escala quanto na outra, para $r = 1,\ldots,5$ tem-se, respectivamente, os percentuais de gordura nas regiões dos braços, pernas, tronco, androide e ginecoide.

TABELA 9 – VARIÁVEIS RESPOSTAS NAS ESCALAS ORIGINAL Y_r E TRANSFORMADA Y_r^* PARA AS SEIS PRIMEIRAS LINHAS DO CONJUNTO DE DADOS

	Escala original					Escala transformada				
Indivíduo	Y_1	Y_2	Y_3	Y_4	Y_5	Y_1^*	Y_2^*	Y_3^*	Y_4^*	Y_5^*
1	0,163	0,234	0,238	0,295	0,314	0,239	0,328	0,363	0,438	0,428
2	0,331	0,335	0,366	0,432	0,432	0,572	0,527	0,634	0,708	0,668
3	0,252	0,312	0,179	0,169	0,354	0,415	0,482	0,237	0,190	0,510
4	0,094	0,172	0,206	0,251	0,272	0,102	0,205	0,295	0,352	0,342
5	0,204	0,304	0,247	0,285	0,342	0,320	0,466	0,382	0,419	0,485
6	0,204	0,331	0,244	0,235	0,426	0,320	0,519	0,375	0,320	0,657

FONTE: O autor (2018).

TABELA 10 – RAZÃO DE CHANCES (RC) E INTERVALOS (IC) COM 95% DE CONFIANÇA ASSOCIADOS AO MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO AJUSTADO AOS DADOS DO PERCENTUAL DE GORDURA CORPORAL

	Braços	Pernas	Tronco	Androide	Ginecoide
Efeito	RC (IC 95%)	RC (IC 95%)	RC (IC 95%)	RC (IC 95%)	RC (IC 95%)
`Idade`	1,004(1,002-1,006)	1,000(0,998-1,002)*	1,004(1,002-1,006)	1,005(1,002-1,007)	0,999(0,997-1,001)*
`IMC`	1,091(1,078-1,104)	1,068(1,055-1,080)	1,104(1,092-1,115)	1,103(1,090-1,117)	1,061(1,051-1,072)
`Sexo-M`	0,411(0,384-0,439)	0,426(0,400-0,455)	0,669(0,632-0,709)	0,709(0,663-0,758)	0,484(0,458-0,511)
`IPAQ-IA`	0,888(0,810-0,974)	0,956(0,872-1,048)*	0,899(0,828-0,977)	0,891(0,810-0,980)	0,936(0,864-1,013)*
`IPAQ-A`	0,780(0,713-0,853)	0,864(0,792-0,944)	0,839(0,775-0,909)	0,837(0,764-0,918)	0,864(0,800-0,932)

FONTE: O autor (2018).

APÊNDICE F – CÓDIGOS EM R USADOS NA ANÁLISE DOS DADOS DO PERCENTUAL DE GORDURA CORPORAL

```
# Data analysis II: Percentage of body fat
# Author: Ricardo Rasmussen Petterle/UFPR
# Loading extra packages
require(mcglm)
require(Matrix)
# Loading extra functions
source("mc_odds_ratio.R")
# Loading data set
data <- read.table("Dados.csv", header = TRUE,
                   sep = ",", na.strings = "")
# Preparing data set
data$ARMS <- data$ARMS/100
data$LEGS <- data$LEGS/100
data$BODY <- data$BODY/100
data$ANDROID <- data$ANDROID/100
data$GYNECOID <- data$GYNECOID/100
data$SEX <- as.factor(data$SEX)
data$IPAQ <- as.factor(data$IPAQ)
# Matrix linear predictor
Z0 <- mc_id(data)
data$Ntrial <- 1
# Univariate models
# Response variable: Arms % of fat
# Linear predictor
form.arms <- c(ARMS ~ AGE + IMC + SEX + IPAQ)
# Fit
fit.arms <- mcglm(linear_pred = form.arms,
                  variance = "binomialP",
                  link = "logit",
                  Ntrial = list(data$Ntrial),
                  matrix_pred = list("resp1" = Z0),
                  data = data,
                  control_algorithm = list(tol = 1e-04,
                                           tuning = 0.8))
```

```
# Response variable: LEGS % of fat
# Linear predictor
form.legs <- c(LEGS ~ AGE + IMC + SEX + IPAQ)
# Fit
fit.legs <- mcglm(linear_pred = form.legs,
                  variance = "binomialP",
                  link = "logit",
                  Ntrial = list(data$Ntrial),
                  matrix_pred = list("resp2" = Z0),
                  data = data,
                  control_algorithm = list(tol = 1e-04,
                                           tuning = 0.8))
# Response variable: BODY % of fat
# Linear predictor
form.body <- c(BODY ~ AGE + IMC + SEX + IPAQ)
# Fit
fit.body <- mcglm(linear_pred = form.body,
                  variance = "binomialP",
                  link = "logit",
                  Ntrial = list(data$Ntrial),
                  matrix_pred = list("resp3" = Z0),
                  data = data,
                  control_algorithm = list(tol = 1e-04,
                                           tuning = 0.8))
# Response variable: ANDROID % of fat
# Linear predictor
form.and <- c(ANDROID ~ AGE + IMC + SEX + IPAQ)
# Fit
fit.and <- mcglm(linear_pred = form.and,
                 variance = "binomialP",
                  link = "logit",
                  Ntrial = list(data$Ntrial),
                  matrix_pred = list("resp4" = Z0),
                  data = data,
                  control_algorithm = list(tol = 1e-04,
                                           tuning = 0.8))
```

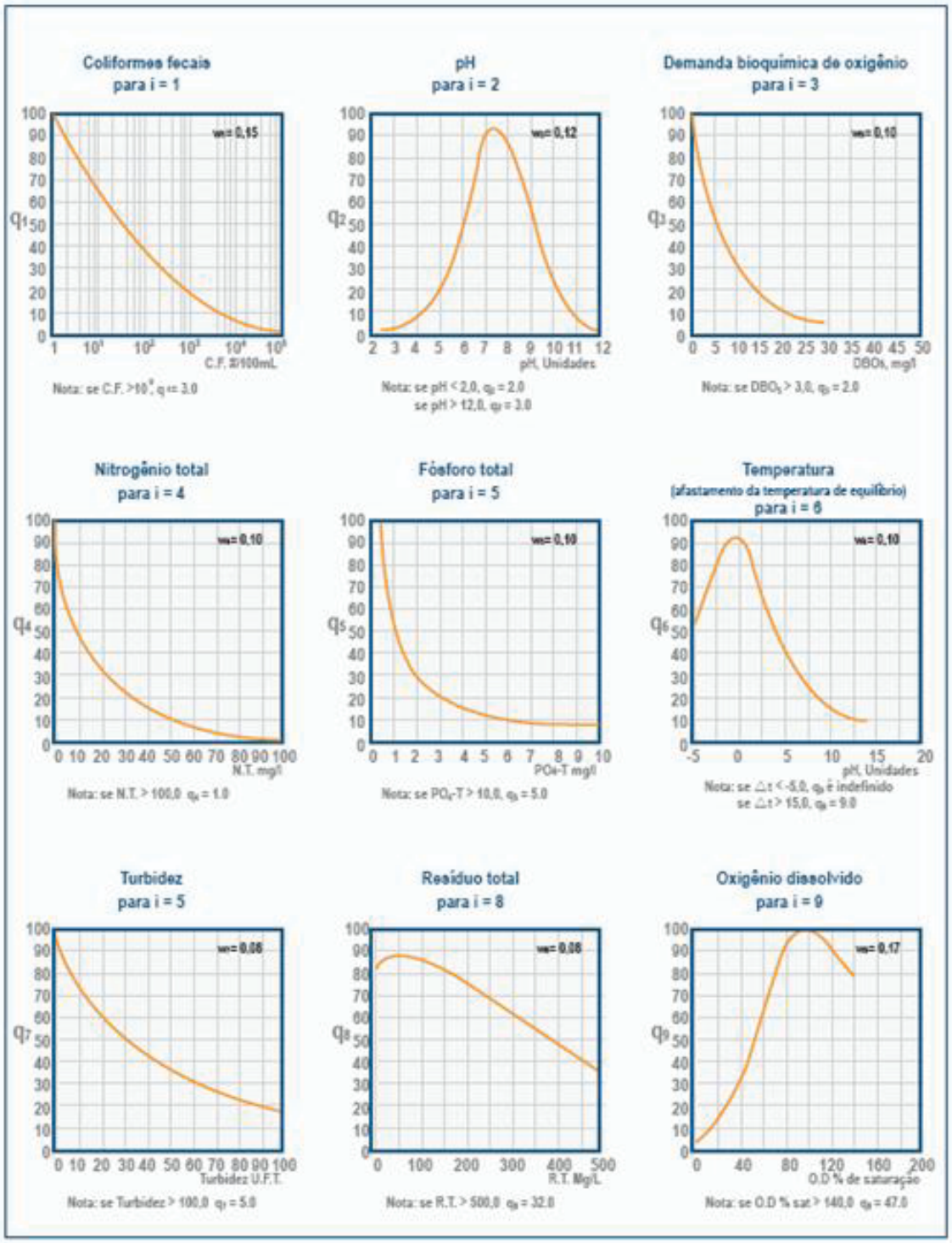
```
# Response variable: GYNECOID % of fat
# Linear predictor
form.gy <- c(GYNECOID ~ AGE + IMC + SEX + IPAQ)
# Fit
fit.gy <- mcglm(linear_pred = form.gy,
                variance = "binomialP",
                   link = "logit",
                   Ntrial = list(data$Ntrial),
                   matrix_pred = list("resp5" = Z0),
                   data = data,
                   control_algorithm = list(tol = 1e-04,
                                            tuning = 0.8))
# Multivariate model
# Linear predictors
form.arms <- c(ARMS ~ AGE + IMC + SEX + IPAQ)
form.legs <- c(LEGS ~ AGE + IMC + SEX + IPAQ)
form.body <- c(BODY ~ AGE + IMC + SEX + IPAQ)
form.and <- c(ANDROID ~ AGE + IMC + SEX + IPAQ)
form.gy <- c(GYNECOID ~ AGE + IMC + SEX + IPAQ)
# Fit
fit.joint <- mcglm(linear_pred = c(form.arms, form.legs, form.body,
                                   form.and, form.gy),
                   matrix_pred = list(Z0, Z0, Z0, Z0, Z0),
                   link = c("logit", "logit", "logit",
                            "logit", "logit"),
                   variance = c("binomialP", "binomialP", "binomialP",
                                "binomialP", "binomialP"),
                   data = data,
                   control_algorithm = list(tol = 1e-04,
                                            tuning = 0.8))
# Comparing uni and multivariate models (goodness-of-fit)
rbind(gof(list(fit.arms, fit.legs, fit.body, fit.and, fit.gy)),
                gof(fit.joint))
# ANOVA Wald test
anova(fit.joint)
# Regression parameters
summary(fit.joint, print = "Regression")
```

```
# Odds ratio and CI 95%
names = c("Idade","IMC","Sexo-M",
          "IPAQ-IA","IPAQ-A")
mc_odds_ratio(fit = fit.joint, names = names, response = 1)
mc_odds_ratio(fit = fit.joint, names = names, response = 2)
mc_odds_ratio(fit = fit.joint, names = names, response = 3)
mc_odds_ratio(fit = fit.joint, names = names, response = 4)
mc_odds_ratio(fit = fit.joint, names = names, response = 5)
# Dispersion parameter
summary(fit.joint, print = "Dispersion")
# Correlation parameter
summary(fit.joint, print = "Correlation")
# END
```

Anexos

ANEXO A – CURVAS DE VARIAÇÃO DE QUALIDADE DA ÁGUA

FIGURA 36 – CURVAS MÉDIAS DE VARIAÇÃO DE QUALIDADE DA ÁGUA



FONTE: AGÊNCIA NACIONAL DE ÁGUAS (2018).

ANEXO B – AUTORIZAÇÃO PARA USO DO CONJUNTO DE DADOS DO PERCENTUAL DE GORDURA CORPORAL.

SEMPR

SERVIÇO DE ENDOCRINOLOGIA E METABOLOGIA
DO HOSPITAL DE CLÍNICAS
DA UNIVERSIDADE FEDERAL DO PARANÁ

DECLARAÇÃO

Declaro, para os devidos fins, que parte dos dados utilizados na dissertação de mestrado denominada Diagnóstico Densitométrico de Sarcopenia, aprovado pelo CEP (CAAE: 16596713.7.0000.0096), foi cedida para o pesquisador Ricardo R. Petterle, como extensão do projeto realizado no Serviço de Endocrinologia e Metabologia do Paraná (SEMPR).

À disposição

Thaisa H. Jonasson
Endocrinologista

Dra Victória Z. C. Borba
Professora Adjunta de Endocrinologia da UFPR Chefe do Serviço de Endocrinologia e Metabologia do HC UFPR(SEMPR)

16/02/2018

Rua: Avenida Agostinho Leão Jr, 285 Alto da Glória, CEP 80030-110 Curitiba Paraná Brasil
Contato: 55 (41) 2141-1730